Monaural and binaural masking release with speech-like stimuli

Hyojin Kim[1]*, Viktorija Ratkute[2], Bastian Epp[1]

[1]Hearing Systems Section, Department of Health Technology, Technical University of Denmark, Kgs. Lyngby, Denmark

[2]Biomedical Engineering, Technical University of Denmark, Kgs. Lyngby, DK-2800, Denmark

*hykim@dtu.dk

ABSTRACT

Comodulated masking noise and binaural cues can facilitate grouping frequency components into the target sound and noise. This results in a decrease in detection thresholds, quantified as comodulation masking release (CMR) and binaural masking level difference (BMLD), respectively. However, their relevance to speech perception is unclear as most studies have used artificial stimuli different from speech. Here, we investigated the ecological validity of masking release by using stimuli with speech-like spectro-temporal dynamics. We designed stimuli reflecting formant changes in speech. Each stimulus consists of three masker bands centered at formant frequencies F1, F2, and F3 based on CV combination: /gu/, /fu/, and /pu/. Maskers were either uncorrelated or comodulated to estimate CMR. In addition, target tone was centered at F2 trajectories and presented either with no IPD or IPD of π to induce BMLD. We found that the CMR was small ($<$ 3 dB) while BMLD was comparable to previous findings ($\sim$ 9 dB). To conclude, we suggest that features other than comodulation play a dominant role in grouping frequency components, such as the spectral proximity, the number of masker bands, and and formant trajectories.

1. Introduction

In a complex acoustic environment, frequency components of multiple sound sources are mixed in time. Nevertheless, the auditory system can separate the frequency components into a target sound and a masking sound. The auditory system uses distinctive features of the target sound to form separate streams, such as "proximity" or "similarity" of frequency components of the target sound [Bregman, 1994]. When frequency components in artificial stimuli are close to each other in frequency or in time, they tend to be grouped into one stream. For example, if two pure tones are being played closely in frequency or time, they are perceived as one stream rather than two (e.g., trill threshold) [Miller and Heise, 1950]. The frequency components of natural sounds are, however, separated too far apart to be grouped by spectral proximity. For instance, speech shows multiple prominent frequency components called formants [Raphael et al., 2007]. The vowel /i/ can have three formants (F1, F2, and F3) at 270 Hz, 2290 Hz, and 3010 Hz, respectively [Raphael et al., 2007]. Whether these frequency components can be grouped into one stream when played continuously is unknown. In addition to the spectral proximity, formants show another signal property that can facilitate stream formation. Formants show coherent amplitude modulation patterns as a consequence of vocal tract vibration [Raphael et al., 2007]. Such comodulation can be a grouping cue by similarity for the auditory system, which can play a role in speech perception [Nelken et al., 1999]. Supporting the ecological relevance of comodulation, previous studies regarding the discrimination of speech sounds showed that dyslexic children showed an insensitivity to the amplitude modulation rather than to formant changes [Ziegler and Goswami, 2005, Goswami et al., 2011].

In psychoacoustical experiments, the benefit of comodulation has been shown as an enhanced target detection performance [Hall et al., 1984]. For instance, the detection threshold of a tone masked by comodulated noise is lower compared to the threshold of a tone in noise with uncorrelated modulation patterns. This difference in masked threshold is commonly referred to as comodulation masking release (CMR) [Hall et al., 1984]. Adding comodulated noise components that are remote from the masked tone in frequency allows the detection threshold to be further decreased, resulting in increased CMR [Hall et al., 1990]. CMR paradigms can be divided into two main groups commonly referred to as "band-widening experiments" and "flanking-band experiments". In the band-widening experiment, a single band of noise is used as a masker centered around the signal frequency. This masker is either modulated or uncorrelated. Masked thresholds are measured as a function of the masker bandwidth [Hall et al., 1984, Carlyon et al., 1989]. Masked threshold decrease with increased masker bandwidth, even when exceeding the critical bandwidth [Hall et al., 1984], indicating that CMR processing occurs across frequency channels. In the flanking-band paradigm, a masker consists of the signal centered band and several narrow bands remotely centered around the target signal with spectral spacing (flanking bands, FBs). From the studies using the flanking-band paradigm, a general agreement is that CMR increases with increased spectral proximity between the target signal and FBs, and the number of FBs [Hall et al., 1990]. Both experiment paradigms showed the large benefit of comodulation in target detection, and the effective frequency range of CMR processing. However, the frequency settings of CMR stimuli are very dissimilar to speech. Compared to the static frequency components in most CMR experiments, each phoneme in speech has specific formant settings in frequency and shows formant transition from one phoneme to another. Few CMR studies used sweeping stimuli with a certain rate or frequency components with random center frequencies fluctuations [Verhey et al., 2013, Doleschal and Verhey, 2020]. Still, these stimuli are not similar to formant changes in speech.

In addition, spatial information can induce masking release binaural masking level difference (BMLD). The ecological validity of binaural cues has been shown with speech signals where BMLD or BILD were ranging from 5 - 12 dB [Ho et al., 2015, Johansson and Arlinger, 2002]. A previous study showed that comodulation and IPD showed a superposition of masking release [Epp and Verhey, 2009]. In this study, we aim to investigate the interplay of comodulation, formant transitions, and spatial cues in a masking release experiment. Here, we assumed that the fundamental unit of speech may be a syllable rather than a phoneme based on previous studies by [Goswami et al., 2011, Vazeux et al., 2020]. We hypothesized that a syllable is the result of the natural grouping of frequency components by comodulation. We analyzed speech samples of consonant and vowel combinations (CV). We extracted formant patterns and created three masker bands at formant frequencies F1, F2, and F3 based on CV combination: /gu/, /fu/, and /pu/. Each resulting CV sound had different degrees of formant changes: i) a stationary formant transition (/gu/), ii) a gradual and correlated formant transition across three formants (/fu/), and iii) a highly variable and uncorrelated formant transition across three formants (/pu/). Furthermore, we investigated whether spatial information interaural phase difference (IPD) would act as a grouping cue in speech-like stimuli. We measured BMLD by adding an IPD of π to the target tone. Under the assumption that formant transitions are a common property of everyday sound,

we hypothesized that speech-like stimuli can induce CMR and BMLD regardless of the degree of formant changes comparable to previous findings with a fixed flanking-band paradigm [e.g. Epp and Verhey, 2009].

2. Methods

2.1. Stimuli

We designed stimuli that have formant patterns extracted from speech samples /gu/, /fu/, and /pu/ (Figure 1). Speech samples were extracted from the Oldenburg Logatome Corpus (OLLO) speech database [Wesker et al., 2005]. We chose VCV combination syllables with German speakers with no dialect. We chose one speaker (female) for analysis in order to preserve the nature of the speech signal. Due to the differences in formant transitions across speakers, averaging sound samples would not reflect natural formant transitions. We extracted CV combinations with a vowel detection algorithm to find the onset and endpoint of each phoneme [Garnaik et al., 2020]. To extract F1, F2 and F3 formants from the sound samples, the PRAAT software package was used [Boersma, 2011]. Five formants by frame were extracted with the formant ceiling of 5500 Hz. The duration of the analysis window length was 0.01 seconds. The first three formants were chosen as they are relevant for speech perception [Raphael et al., 2007]. Following the trajectories of the formants, we placed masker bands where each center frequency follows formant patterns (F1, F2, and F3) of /gu/, /fu/, and /pu/. The target signal was centered at F2. Each masker band had a bandwidth of 20 Hz. The duration of the signal was 500 ms, including 20 ms raised-cosine ramps at the masker onset and offset.

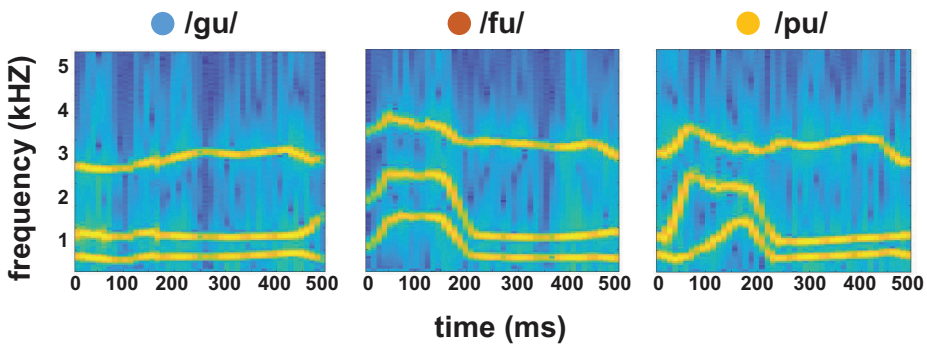

Figure 1: Spectrograms of the stimuli: /gu/, /fu/, and /pu/ from left to right.

All the signals were generated digitally at a sampling rate of 44.1 kHz in MATLAB 2018b (TheMathworks, Natick, MA). The instantaneous frequency extracted from the formant analysis with PRAAT was used as a carrier and multiplied with a low-pass-filtered noise in order to generate narrow-band noise masker components m_i:

$$m_i(t) = a_i \cdot sin\left(\int_0^t 2\pi \cdot f_i(t')dt' + \phi_i\right) \cdot N_i(t) \tag{1}$$

where N_i is low-pass-filtered noise without a DC component, a_i is amplitude, and it was chosen such that the overall level of the masker was 60 dB. ϕ_i represents the starting phase of the carrier, chosen randomly for each band. f_i is the time-varying frequency estimated from the formant analysis. For the uncorrelated noise, each masker band was multiplied with each time-varying frequency band (formant). For the comodulated noise, the same modulator was used for all masker bands. The target tone was added, which followed the center frequency of the middle masker band:

$$s(t) = a_s \cdot sin\left(\int_0^t 2\pi f_s(t')dt'\right) \tag{2}$$

where a_s is the amplitude and f_s is the signal frequency at each time point.

2.2. Protocol

After the training session, each listener performed three threshold measurements for all conditions. For each measurement, conditions were presented in randomized order. The thresholds were estimated by averaging three trials. One additional measurement was done if the thresholds from the last three measurements had a high standard deviation (SD > 3 dB). In case of high SD between 3 trials, the measurement of that condition was repeated. Although, after repetition, four participants had an SD of 4 dB in one condition and one participant in two conditions. During the threshold measurement, listeners were seated in a double-walled, soundproof booth with ER-2 headphones. We used an adaptive, three-interval, three-alternative forced-choice procedure (3-AFC) with a one-up, two-down rule to estimate the 70.7% of the psychometric function [Ewert, 2013], [Levitt, 1971]. Three sound intervals were presented with a pause of 500 ms in between. Two intervals contained the maskers only. The remaining interval contained the target tone with maskers. The listeners' task was to choose the interval with the target tone by pressing the corresponding number key (1, 2, 3) on the keyboard. Visual feedback was provided, indicating whether the answer was "WRONG" or "CORRECT." The target tone's start level was set to 75 dB, and the tone level was adjusted with the initial step size of 8 dB. The step size was halved after each lower reversal until it reached the minimum step size of 1 dB. The signal level at a minimum step size of 1dB was measured six times, and the mean of those measurements was estimated as the threshold.

2.3. Listeners

From hearing screening, we recruited ten normal-hearing listeners. None of them reported any history of hearing impairment and had pure-tone hearing thresholds within 15 dB HL for the standard audiometric frequencies from 125 to 4000 Hz. All participants provided informed consent, and all experiments were approved by the Science-Ethics Committee for the Capital Region of Denmark (reference H-16036391).

3. Results

We measured both CMR and BMLD with speech-like stimuli. Results are shown in Figure 2. The mean CMR for each condition was: /gu/ (1.03 dB), /fu/ (0.84 dB), /pu/ (0.11 dB) in diotic the conditions, and /gu/ (2.4 dB), /fu/ (1.83 dB), and /pu/ (1.22 dB) in the dichotic conditions. No significant difference was found across all CV combinations in both diotic and dichotic conditions (one-way ANOVA, in diotic condition: F(6.3) = 0.82, p = 0.54). Although the mean values of CMR across conditions showed no significant difference, CMR values showed a high variability in /fu/ and /pu/ maskers (Figure 2 with grey scatter plot). For /fu/with IPD of 0, CMR values were from - 6.11 dB to 5.28 dB. For /pu/ with IPD of π, CMR values showed high variability ranging from - 6.8 dB to 10.3 dB (individual data is available in Supplement).

The mean BMLD for each condition was: /gu/ (8.46 dB), /fu/ (8 dB), /pu/ (9.83 dB) in diotic conditions, and /gu/ (9.01 dB), /fu/ (10.02 dB), and /pu/ (9.04 dB) in dichotic conditions. No significant difference was found across all CV combinations in both diotic and dichotic conditions (one way ANOVA, F(6.02) = 0.25, p = 0.94).

4. Discussion

In this study, we investigated the ecological validity of masking release, CMR and BMLD. For BMLD, the results were in line with the study with voiced vowels where BMLD was ranged from 7.3 to 11.5 dB [Ho et al., 2015], and with speech signals where the binaural intelligibility level difference (BILD) ranged from 5.7 to 7.7 dB [Johansson and Arlinger, 2002]. However, CMR values were not significant compared to previous CMR studies with artificial stimuli. In the previous study by Verhey et al. [2013], they measured CMR with sweeping maskers. Proportional to the increase in sweep rate, the CMR was reduced (e.g., from 11 dB to 6 dB). In another CMR study with random-variations in frequencies of maskers, CMR was small (e.g., 5 dB) compared to a CMR study with stationary stimuli (e.g., 6 – 10 dB) [Doleschal and Verhey, 2020]. The stimuli used in this experiment had different degrees of formant variation (stationary, correlated formant transitions, uncorrelated formant transitions). For the condition with stationary formant changes (e.g., /gu/), CMR was less than 3 dB while CMR showed individual variability ranging from - 6 dB to 10 dB with more formant variations (e.g., /fu/, /pu/).

Based on previous studies by Hall et al. [1990], the number of FBs may play a role in facilitating the grouping of formants by comodulation. Here, one of the differences in the stimulus design was the number of flanking bands and the spectral proximity between flanking bands. While the study by Verhey et al. [2013] with sweeping masker had five masker bands, the stimuli used in this study had three FBs and a wider spacing than the sweeping masker. This factor may be related to the reduced amount of CMR in our study. Another

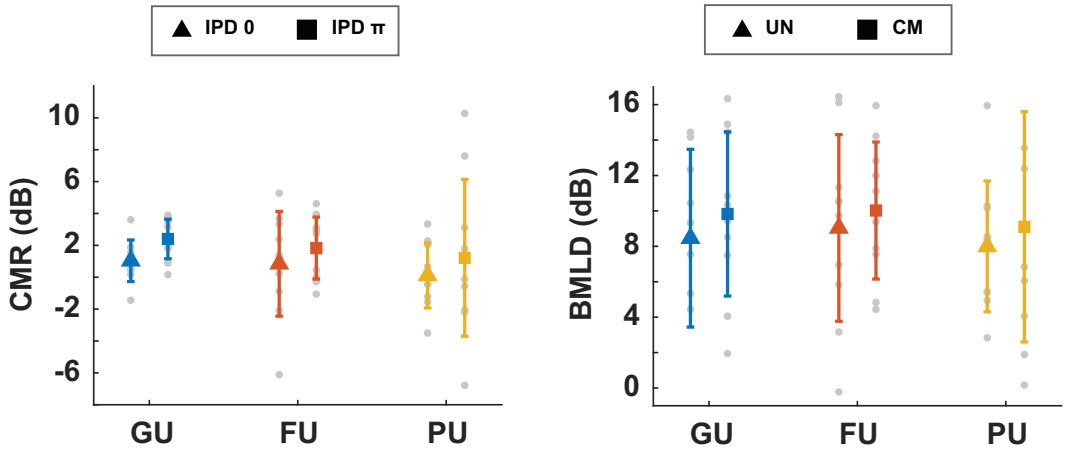

Figure 2: The plots of CMR and BMLD. The mean CMR for each condition is plotted on the left. The mean BMLD for each condition is plotted on the right. Blue color represents /gu/, orange color represents /fu/, and yellow color represents /pu/. Circle indicates conditions with IPD of 0 and filled squired indicates conditions with IPD of π. Error bars indicate plus-minus one standard deviation.

difference in stimuli was non-stationary frequency changes with formant patterns (e.g., /fu/ and /pu/). While previous studies used either randomly varied frequency in time or the same sweeping patterns across noise bands, we extracted formant changes from speech samples. With more formant variations, some listeners showed higher positive CMR or negative CMR than the stimuli with stationary formant patterns (e.g., /gu/). In the condition with /fu/ with IPD of 0, the maximum value of CMR was 5.28 dB while the minimum value of CMR was - 6.11 dB. Similarly, in the condition with /pu/ with IPD of π, the maximum value of CMR was 10.28 dB, while the minimum value of CMR was - 6.78 dB. We analyzed the relationship between CMR and BMLD with linear regression (Figure 3). The /pu/ in dichotic conditions showed a positive correlation between CMR and BMLD. That is, the enhancement of CMR and BMLD occur simultaneously. We speculate that the IPD cue facilitated the segregation of the target tone from uncorrelated formant transition of /pu/, which in turn enhanced the grouping of maskers to use comodulation cue. Similarly, in /fu/ conditions, both enhanced and reduced CMR were observed. Unlike /pu/ condition, /fu/ conditions did not show a significant relationship between CMR and BMLD. This may suggest that the correlated formant transition either facilitated the stream formation across formants or served as a distracting cue to detect the target tone. Thus, the underlying neural basis for CMR and BMLD interaction needs to be further investigated to explain how formant transitions affect the use of IPD cues differently in /pu/ and /fu/ conditions.

	$y = a * x + b$		
diotic	*a*	*b*	*p-values*
/gu/	1.85	6.55	0.16
/fu/	-0.67	8.56	0.07
/pu/	-1.14	9.95	0.14
dichotic			
/gu/	1.03	6.62	0.58
/fu/	-0.48	10.89	0.51
/pu/	0.71	8.17	0.04*

5. Conclusion

In this study, we evaluated the ecological validity of CMR and BMLD. Our hypothesis was that if CMR and BMLD are relevant to speech perception, speech-like stimuli will induce a comparable amount of CMR and BMLD to

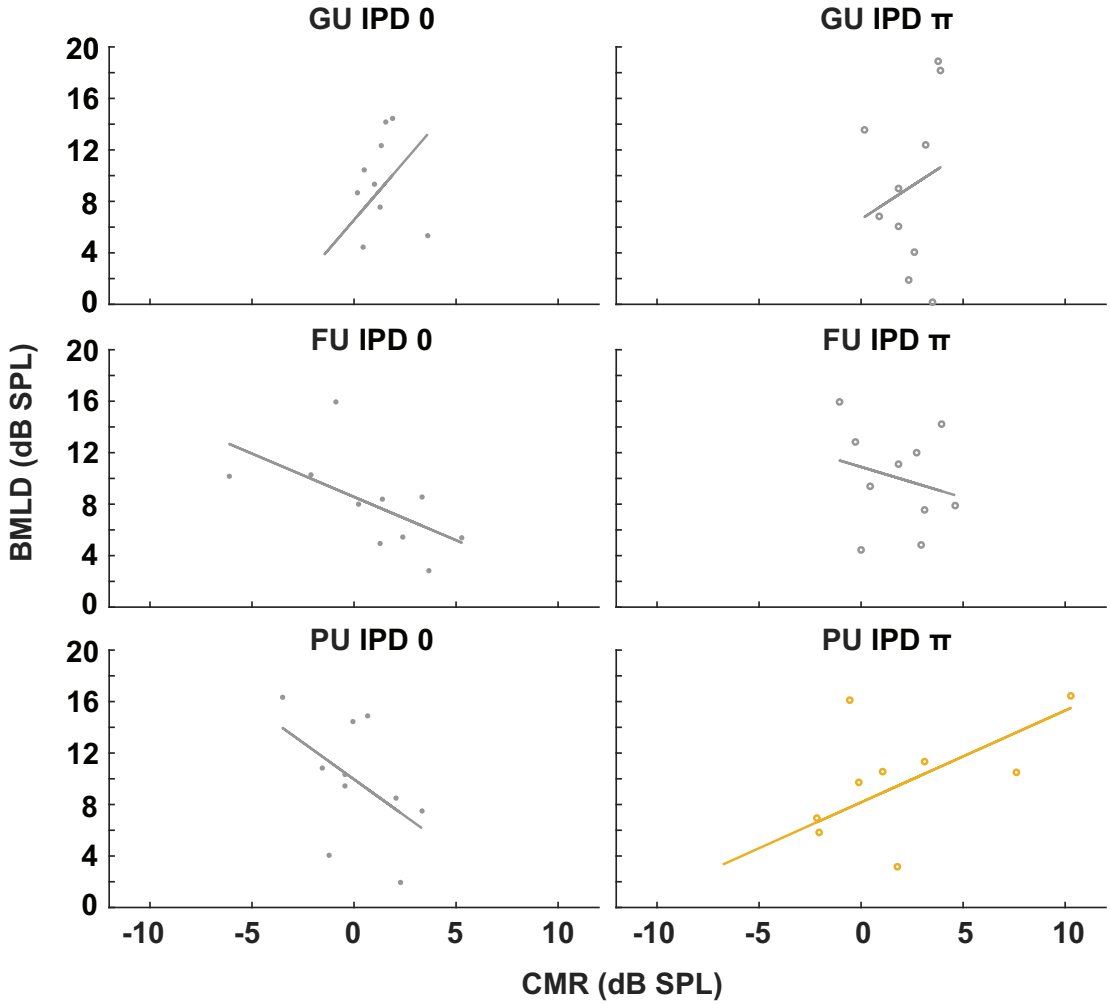

Figure 3: Linear regression analysis between CMR and BMLD for all conditions. Only the conditions with significant p-values were plotted with color (see Table ??)

previous findings. Results showed that CMR was not significant while BMLD was around 9 dB. We concluded that comodulation is not enough cue to group three formants. We suggest that other sound properties such as formant transition patterns, the proximity of masker bands to the target signal, and the number of masker bands can be more relevant for grouping frequency components to induce speech perception.

References

P. Boersma. Praat: doing phonetics by computer [computer program]. *http://www. praat. org/*, 2011.

A. S. Bregman. *Auditory scene analysis: The perceptual organization of sound*. MIT press, 1994.

R. P. Carlyon, S. Buus, and M. Florentine. Comodulation masking release for three types of modulator as a function of modulation rate. *Hearing research*, 42(1):37–45, 1989.

F. Doleschal and J. L. Verhey. Comodulation masking release with random variations of flanking-band center frequencies. *The Journal of the Acoustical Society of America*, 148(2):692–700, 2020.

B. Epp and J. L. Verhey. Superposition of masking releases. *Journal of computational neuroscience*, 26(3): 393–407, 2009.

S. D. Ewert. Afc—a modular framework for running psychoacoustic experiments and computational

perception models. In *Proceedings of the international conference on acoustics AIA-DAGA*, pages 1326–1329, 2013.

S. Garnaik, A. Kumar, G. Pradhan, and K. Sethi. An efficient approach for detecting vowel onset and offset points in speech signal. *International Journal of Speech Technology*, 23(3):643–651, 2020.

U. Goswami, T. Fosker, M. Huss, N. Mead, and D. Szűcs. Rise time and formant transition duration in the discrimination of speech sounds: the ba–wa distinction in developmental dyslexia. *Developmental science*, 14(1):34–43, 2011.

J. W. Hall, M. P. Haggard, and M. A. Fernandes. Detection in noise by spectro-temporal pattern analysis. *The Journal of the Acoustical Society of America*, 76(1):50–56, 1984.

J. W. Hall, J. H. Grose, and M. P. Haggard. Effects of flanking band proximity, number, and modulation pattern on comodulation masking release. *The Journal of the Acoustical Society of America*, 87(1):269–283, 1990.

C.-Y. Ho, P.-C. Li, Y.-C. Chiang, S.-T. Young, and W.-C. Chu. The binaural masking-level difference of mandarin tone detection and the binaural intelligibility-level difference of mandarin tone recognition in the presence of speech-spectrum noise. *PloS one*, 10(4):e0120977, 2015.

M. S. Johansson and S. D. Arlinger. Binaural masking level difference for speech signals in noise: diferencia en el nivel de enmascaramiento binaural para senates vocales en ruido. *International journal of audiology*, 41(5):279–284, 2002.

H. Levitt. Transformed up-down methods in psychoacoustics. *The Journal of the Acoustical society of America*, 49(2B):467–477, 1971.

G. A. Miller and G. A. Heise. The trill threshold. *The Journal of the Acoustical Society of America*, 22(5): 637–638, 1950.

I. Nelken, Y. Rotman, and O. B. Yosef. Responses of auditory-cortex neurons to structural features of natural sounds. *Nature*, 397(6715):154–157, 1999.

L. J. Raphael, G. J. Borden, and K. S. Harris. *Speech science primer: Physiology, acoustics, and perception of speech*. Lippincott Williams & Wilkins, 2007.

M. Vazeux, N. Doignon-Camus, M.-L. Bosse, G. Mahé, T. Guo, and D. Zagar. Syllable-first rather than letter-first to improve phonemic awareness. *Scientific reports*, 10(1):1–12, 2020.

J. L. Verhey, H. Klein-Hennig, and B. Epp. Masking release for sweeping masker components with correlated envelopes. *Journal of the Association for Research in Otolaryngology*, 14(1):139–147, 2013.

T. Wesker, B. Meyer, K. Wagener, J. Anemüller, A. Mertins, and B. Kollmeier. Oldenburg logatome speech corpus (ollo) for speech recognition experiments with humans and machines. In *Ninth European Conference on Speech Communication and Technology*. Citeseer, 2005.

J. C. Ziegler and U. Goswami. Reading acquisition, developmental dyslexia, and skilled reading across languages: a psycholinguistic grain size theory. *Psychological bulletin*, 131(1):3, 2005.

6. Supplement

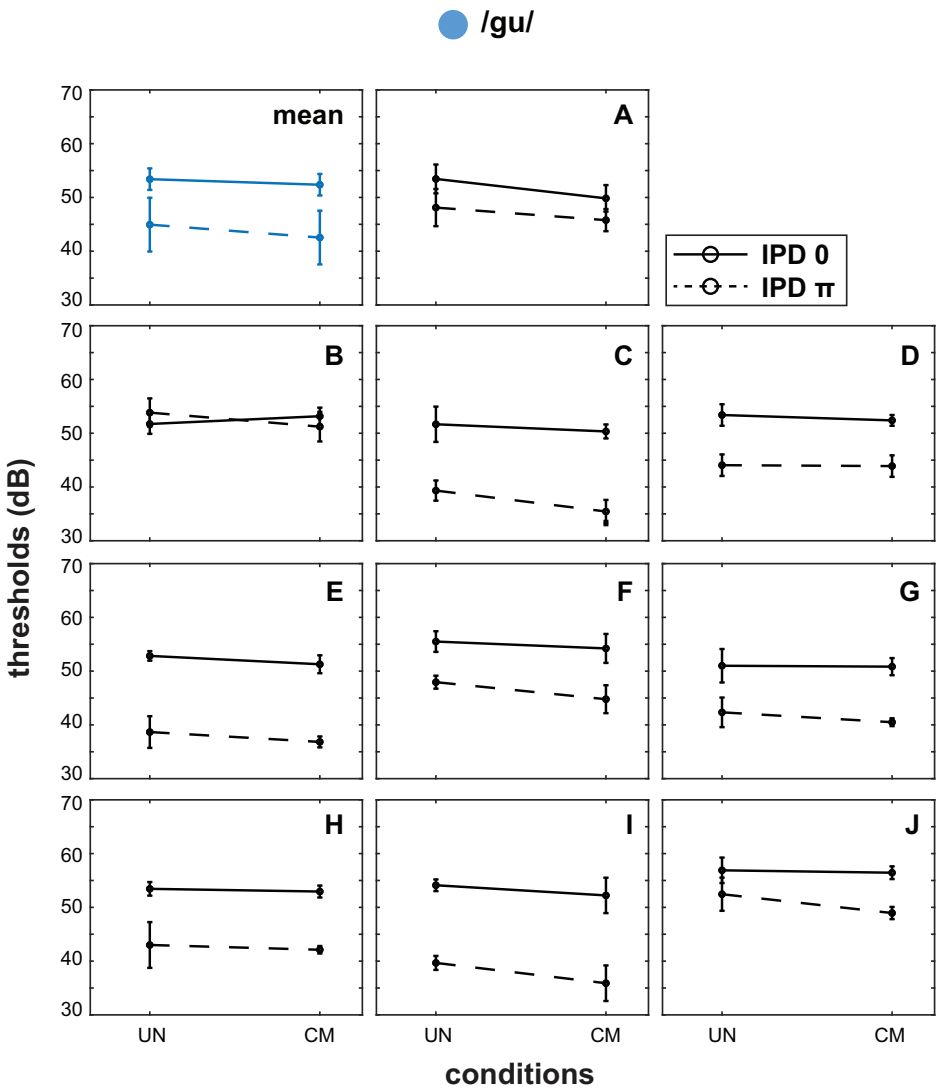

Figure 4: The plots of CMR and BMLD. The mean CMR for each condition is plotted on the left. The mean BMLD for each condition is plotted on the right. Blue color represents /gu/, orange color represents /fu/, and yellow color represents /pu/. Circle indicates conditions with IPD of 0 and filled squired indicates conditions with IPD of π. Error bars indicate plus-minus one standard deviation.

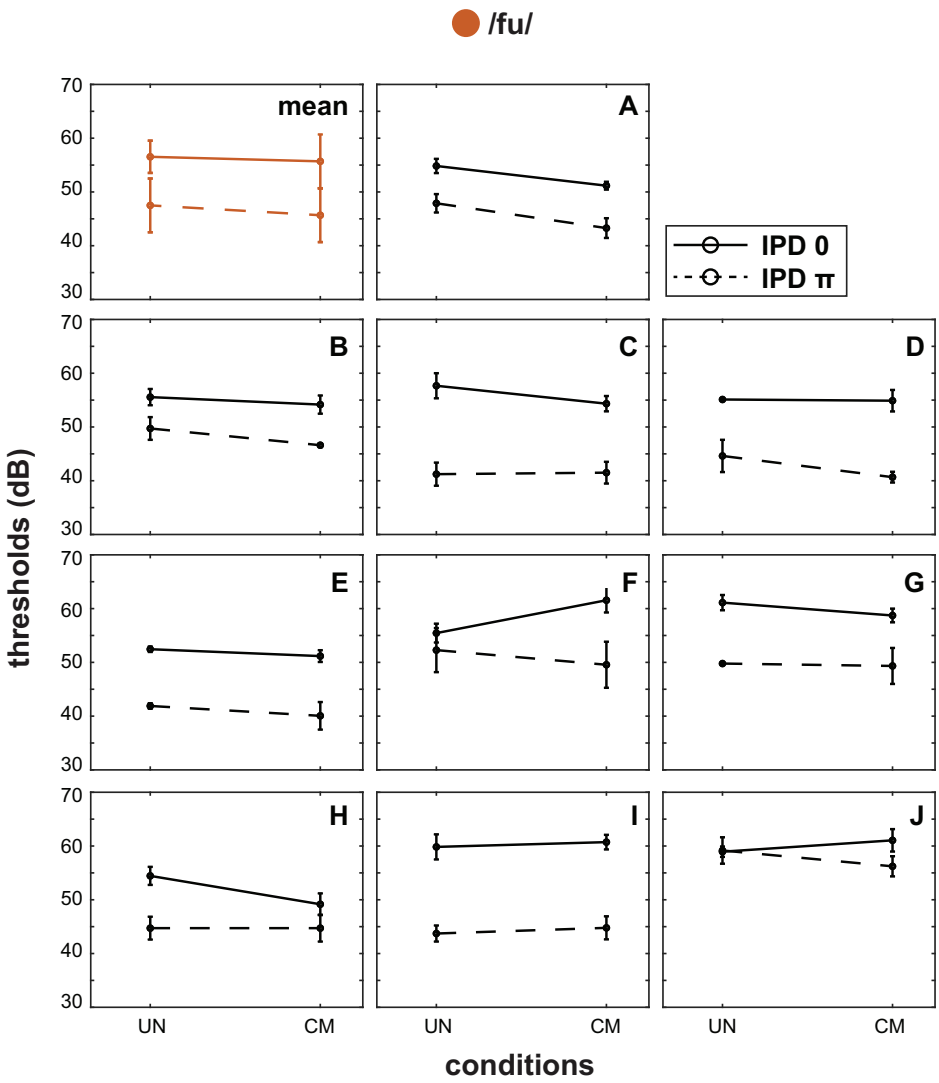

Figure 5: The plots of CMR and BMLD. The mean CMR for each condition is plotted on the left. The mean BMLD for each condition is plotted on the right. Blue color represents /gu/, orange color represents /fu/, and yellow color represents /pu/. Circle indicates conditions with IPD of 0 and filled squired indicates conditions with IPD of π. Error bars indicate plus-minus one standard deviation.

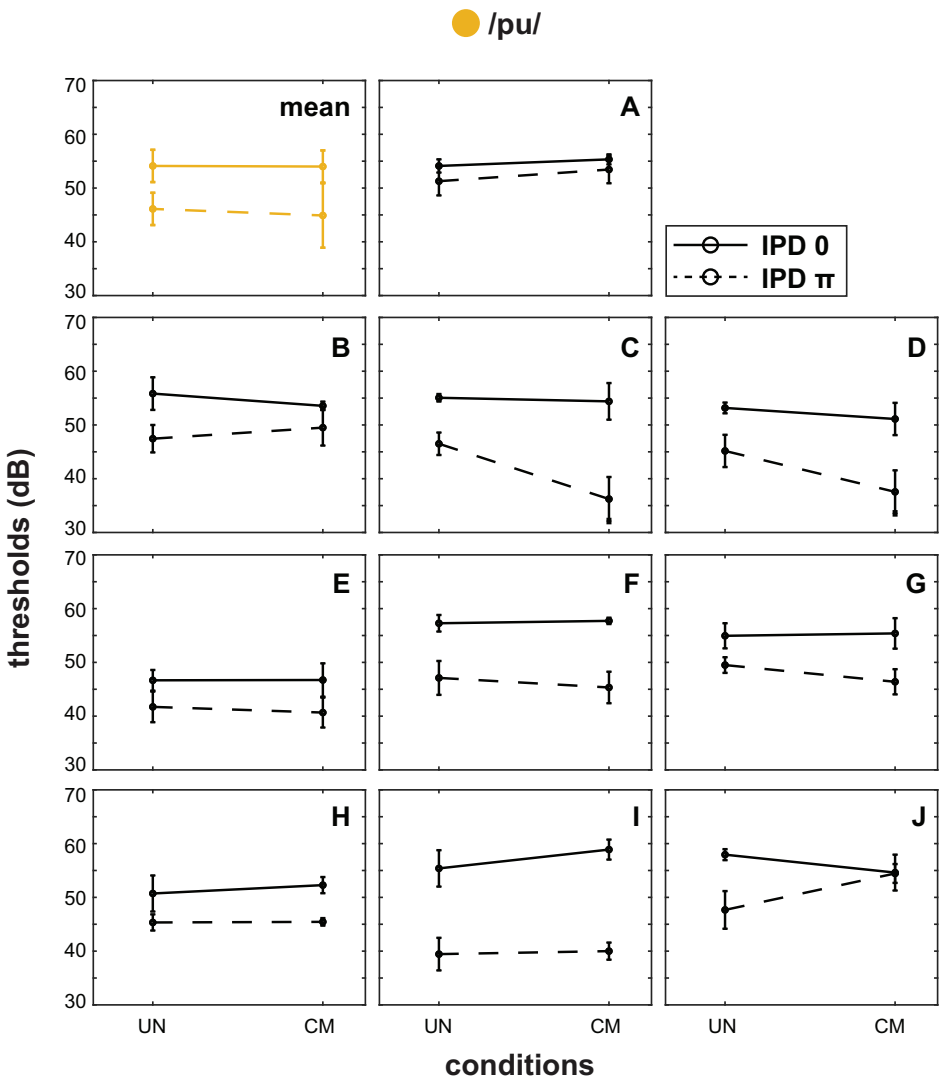

Figure 6: The plots of CMR and BMLD. The mean CMR for each condition is plotted on the left. The mean BMLD for each condition is plotted on the right. Blue color represents /gu/, orange color represents /fu/, and yellow color represents /pu/. Circle indicates conditions with IPD of 0 and filled squired indicates conditions with IPD of π. Error bars indicate plus-minus one standard deviation.