Ecological validity of 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 detecting a target sound from noise. These cues can induce 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 their ecological validity using sounds with speech-like spectro-temporal dynamics. We evaluated the ecological validity of such grouping effect with stimuli reflecting formant changes in speech. We set three masker bands at formant frequencies F1, F2, and F3 based on CV combination: /gu/, /fu/, and /pu/. We found that the CMR was little ($<$ 3 dB) while BMLD was comparable to previous findings ($\sim$ 9 dB). In conclusion, we suggest that other features may play a role in facilitating frequency grouping by comodulation such as the spectral proximity and the number of masker bands.

Introduction

In a complex acoustic environment, frequency components of multiple sound sources are mixed in time. The auditory system can segregate the target sound from noisy backgrounds. The auditory system uses distinct auditory features of the target sound to form a separate stream from other noisy streams. Such features convey "proximity" or "similarity" of frequency components of the target sound [1]. When frequency components 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, they are perceived as one stream rather than two (e.g., trill threshold) [2]. However, in natural sounds such as speech, frequency components are separated far apart to be grouped by spectral proximity. Speech can have multiple prominent frequency components called formants. For example, vowel /i/ can have three formants (F1, F2, and F3) at 270 Hz, 2290 Hz, and 3010 Hz, respectively [3]. With spectral proximity, these formants would not be grouped into the same stream to form a phoneme perception. Instead of spectral proximity, formants show coherent amplitude modulation patterns. As our vocal tract vibrates, it modulates frequency components arising from its vibration. Such co-modulation can be a grouping cue by similarity for our auditory system, which can play a role in speech perception [4]. Furthermore, previous studies in phonetic awareness of children in the pre-reading phase have found that the syllable-first approach improves phonetic awareness compared to the letter-first approach [5]. In addition, in the studies with dyslexia children, dyslexia children showed insensitivity to amplitude modulation rather than to formant frequency changes [6]. Based on these findings, a syllable could be a basic unit of speech as a syllable could be easily connected to learned knowledge.

In psychoacoustical experiments, the benefit of comodulation has been shown as an enhanced target detection performance [7]. For instance, when the target is in comodulated noise, the masked threshold is lower than that in random noise. This decreased masked threshold is quantified as comodulation masking release (CMR). 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 [8]. CMR experiments have used two main methods: "band-widening" and "flanking-bands paradigm." 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 unmodulated, and masked thresholds are measured as a function of the masker bandwidth [7], [9]. Masked threshold decreased with increased bandwidth wider than the critical bandwidth [7]. In the flanking-bands paradigm, a masker consists of several narrow bands. While there is a general agreement where less CMR arises farther from the center band, there is no clear relation between frequency spacing and CMR [8]. Furthermore, frequency settings are artificial and different from natural sounds. Formants change in speech while most CMR experiments use the stimuli with temporally fixed frequency components. Few studies used sweeping stimuli with a certain rate or frequency components with random fluctuations [10], [11]. These patterns are not similar to formant changes in speech. Each phoneme has specific formant settings, and there are formant changes in the transition from one phoneme to another. Here, we aim to investigate its ecological validity with speech-like stimuli. With an assumption that the syllable is a result of natural grouping by modulation, we analyzed speech samples of consonant and vowel combinations (CV). We extracted formant patterns and set three masker bands at formant frequencies F1, F2, and F3 based on CV combination: /gu/, /fu/, and /pu/. Each CV sound has different degrees of formant changes. For example, /gu/ has little formant change, /fu/ has a gradual transition, and /pu/ has high variations in the formant pattern. Our hypothesis was that speech-like stimuli would induce CMR that is comparable to previous findings with a fixed flaking-band paradigm (e.g. 10 dB) [12]. Furthermore, we measured BMLD by adding an interaural phase difference (IPD) to the target tone.

Methods

Stimuli We designed stimuli that have formant patterns extracted from speech samples (Figure 1). We set three masker bands where each center frequency of masker bands follows the formant patterns (F1, F2, and F3) of /gu/, /fu/, and /pu/. The target signal was centered at F2. We set the bandwidth to 20 Hz, and the duration was set to 500 ms, including 50 ms raised-cosine ramps at the masker onset and offset. All the signals were generated digitally at a sampling rate of 44.1 kHz. 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 - starting phase of the carrier and it was chosen randomly for each band. f_i is time-varying frequency estimated from the formant analysis. Center frequency was dynamic across each time point. For the uncorrelated noise, each masker band was multiplied with each time-varying frequency band (formant), and for the comodulated noise, the same modulator was used for all masker bands. Masker level was set to 60 dB. 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.

Protocol After the training session, each listener performed three threshold measurements for all conditions. For each measurement, conditions were in randomized order. The thresholds were estimated by averaging three trials. The additional measurement was done if the thresholds from the last three measurements had high variance (SD > 3 dB). In case of high variance between 3 trials (SD $>$ 3 dB), the measurement of specific conditions was repeated. Although, after repetition, four participants had SD of 4 dB in one condition, and one participants in two conditions. During the threshold measurement, the 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 [13], [14]. Three sound intervals were presented with a pause of 500 ms in between. Two were with only maskers, and the one was the target tone with maskers. The listeners' task was to choose the interval with a target tone by pressing the corresponding number key (1, 2, 3) on the keyboard. Whenever the listener pressed 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 depending on the answer, 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.

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. One listener was tested with 20 dB at 125 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).

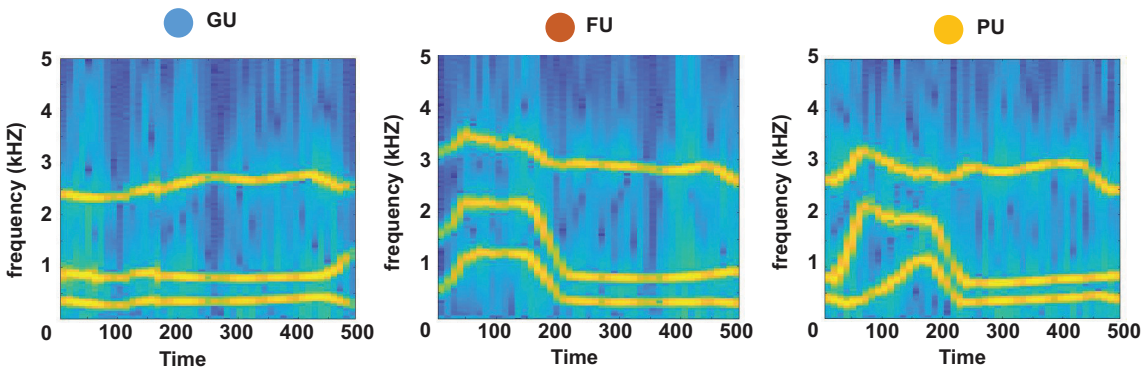

Figure 1: Spectrograms of the stimuli: /gu/, /fu/, and /pu/ from left to right.

Results

We measured both CMR and BMLD with speech-like stimuli. Results are shown in Figure 2. The mean CMR for each conditoin was: /gu/ (1.03 dB), /fu/ (0.84 dB), /pu/ (0.11 dB) in diotic conditions, and /gu/ (2.4 dB), /fu/ (1.83 dB), and /pu/ (1.22 dB) in dichotic conditions. No significant difference was found between the stimulus (one way ANOVA). For conditions with IPD of 0, F(2,37) = 0.43, p = 0.65, and for conditions with IPD of π, F(3.50) = 0.36, p = 0.70.

The mean BMLD for each conditoin 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

between the stimulus (one way ANOVA). For conditions with the uncorrelated masker, $F(2,74) = 0.12$, $p = 0.88$, and for conditions with comodulated masker, $F(2,36) = 0.09$, $p = 0.91$.

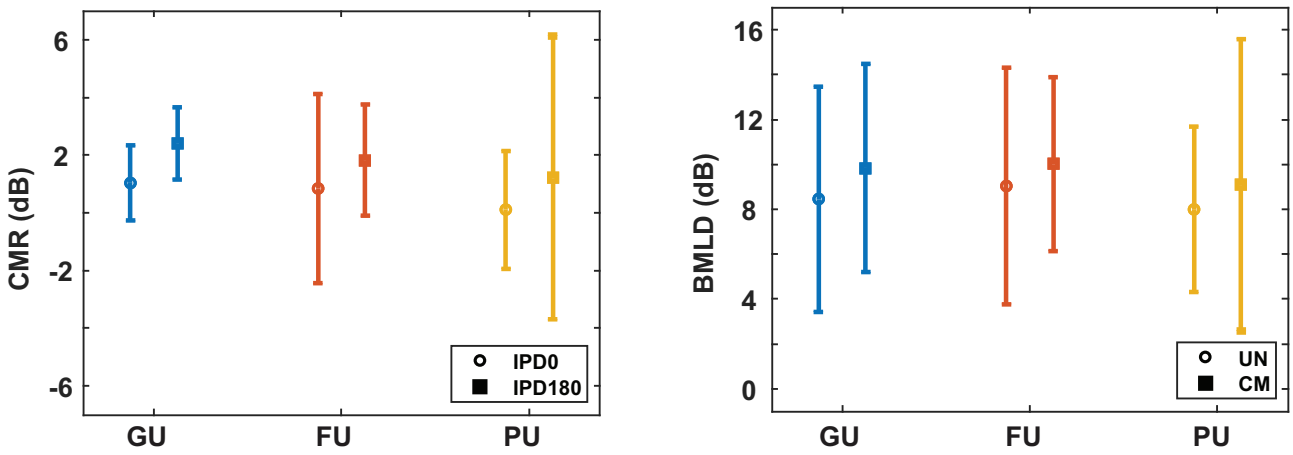

Figure 2: The plots of CMR and BMLD. The mean CMR for each condition is plotted on the left. The mean BMLD for each condition is plotted on the right. Blue color represents /gu/, orange color represents /fu/, and yellow color represents /pu/. Circle indicates conditions with IPD of 0 and filled squired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

Discussion

In this study, we investigated the ecological validity of masking release, CMR and BMLD. In the previous study by Verhey et al. [10], 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). The stimuli used in this experiment had different degrees of the formant variation (stationary, little fluctuation within one phoneme, high fluctuation within one phoneme). For the condition with rather stationary formant changes (e.g. /gu/), CMR was less than 3 dB. One of the difference in the stimulus design was the number of flanking bands and the spectral proximity between flanking bands. While the study with sweeping masker had five masker bands, the stimuli used in this study had three flaking bands, and had wider spacing than the sweeping masker. Based on previous studies with varying number of flanking bands and frequency spacing between flanking bands [8], these features may play a role in facilitating the grouping of formants by comodulation. In another CMR study with random-variations in frequencies of maskers, CMR was small (e.g., 5 dB) compared to other CMR study with stationary stimuli (e.g., 6 – 10 dB) [11]. With non-stationary frequency changes with formant patterns, the stimuli used in this study could not induce CMR comparable to their study. In addition, the stimuli with high variance in formant changes (e.g., /fu/ and /pu/), some listeners showed negative CMR. This may be because the these stimuli sounded more distracting than rather stationary stimulus (e.g. /gu/). This might have made the task of target detection challenging for some listeners. For BMLD, the results were in line the study with voiced vowels where BMLD was ranged from 7.3 to 11.5 dB [15], and with speech signals where the binaural intelligibility level difference (BILD) ranged from 5.7 to 7.7 dB [16].

Conclusion

We evaluated the ecological validity of such grouping effect with stimuli reflecting formant changes in speech. We used six stimuli with three CV combination (e.g., /gu/, /fu/, and /pu/) and spatial cue (e.g. IPD). We found that the CMR was little ($<$ 3 dB) while BMLD was comparable to previous findings ($\sim$ 9 dB). We suggest that other features may play a role in facilitating frequency grouping by comodulation.

References

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

[2] G. A. Miller and G. A. Heise, “The trill threshold,” *The Journal of the Acoustical Society of America*, vol. 22, no. 5, pp. 637–638, 1950.

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

[4] I. Nelken, Y. Rotman, and O. B. Yosef, "Responses of auditory-cortex neurons to structural features of natural sounds," *Nature*, vol. 397, no. 6715, pp. 154–157, 1999.

[5] 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*, vol. 10, no. 1, pp. 1–12, 2020.

[6] 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*, vol. 14, no. 1, pp. 34–43, 2011.

[7] J. W. Hall, M. P. Haggard, and M. A. Fernandes, "Detection in noise by spectro-temporal pattern analysis," *Journal of the Acoustical Society of America*, vol. 76, no. 1, pp. 50–56, 1984.

[8] J. W. Hall III, 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*, vol. 87, no. 1, pp. 269–283, 1990.

[9] R. P. Carlyon, S. Buus, and M. Florentine, "Comodulation masking release for three types of modulator as a function of modulation rate," *Hearing research*, vol. 42, no. 1, pp. 37–45, 1989.

[10] 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*, vol. 14, no. 1, pp. 139–147, 2013.

[11] 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*, vol. 148, no. 2, pp. 692–700, 2020.

[12] B. Epp and J. L. Verhey, "Superposition of masking releases," *Journal of computational neuroscience*, vol. 26, no. 3, pp. 393–407, 2009.

[13] 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*, pp. 1326–1329, 2013.

[14] H. Levitt, "Transformed up-down methods in psychoacoustics," *The Journal of the Acoustical society of America*, vol. 49, no. 2B, pp. 467–477, 1971.

[15] 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*, vol. 10, no. 4, p. e0120977, 2015.

[16] 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*, vol. 41, no. 5, pp. 279–284, 2002.