How Object Formation Can Influence Speech Perception

Siddharth Rajaram1,2, Nicholas Kurkij1,3, Frederick Gallun4, Virginia Best5, Barbara Shinn-Cunningham1,2,3

1 Hearing Research Center, Boston University, 2 Department of Cognitive and Neural Systems, Boston University, 3 Department of Biomedical Engineering, Boston University, 4 Portland VA Medical Center, 5 University of Sydney.

0. ABSTRACT

Listeners automatically group sound into auditory objects, and objects that are high in frequency (e.g., the "high") and low in frequency (e.g., the "low") are separated in space, so when presented in a stream, the location of the low band is not interfered with by the high band. This is optimal because the high and low bands are segregated into different objects, but enhanced when they were grouped together. In contrast, we hypothesized that perception of the location of an individual band would be degraded when the two bands were grouped together. This is optimal because the high and low bands are perceived as different objects. Native English-speaking subjects were asked to identify a monosyllabic word that was band-pass filtered into two frequency bands. We hypothesized that the ability to understand the speech would be degraded when the two bands were grouped into different objects, but enhanced when they were grouped together. In contrast, we hypothesized that perception of the location of an individual band would be degraded when the two bands were grouped together, but enhanced when the bands were perceived in different objects.

1. INTRODUCTION

Motivation: Binaural interference - The perceived lateral location of a high-frequency SAM tone is shown to shift towards the center, when presented simultaneously with a low-frequency diotic SAM tone. When the low-frequency distractor is presented repeatedly in a stream, the location of the target is restored (Best et al., JASA 2007).

Hypotheses: - Binaural interference is caused by obligatory perceptual grouping of the high and low tones - The low-frequency distractor groups with the stream, restoring the target's location - Bands of speech group together (Hall et al., JASA 2004), so this phenomenon should persist with speech tokens - Depending on the task, this grouping may enhance or degrade attributes of the object

2. STIMULI AND TASKS

BUIG Corpus
- 14 speakers (7 male, 7 female)
- 40 monosyllabic words (roughly 300-500 ms)

Localization Conditions (ITD of +/- 300 μs)

Identification Conditions

Speech Bands
The low band was centered at 500Hz, and the high at 2500Hz with 100Hz passbands

3. EXPERIMENTAL PROCEDURES

15 Subjects did the training and testing for the identification and localization on separate days, with the order balanced across subjects

On each day, the subject performed three 40-trial blocks of each condition of the task presented in random order

One subject failed the localization training and was rejected
One more was rejected due to poor performance on all the localization conditions (d' score of < 1)

Training
Identification Task
- Subjects were either trained on the identification task in the low and the high conditions only (order balanced)
- Needed ≥ 75% correct in one 40-trial training block of both the low and the high conditions to pass

Localization Task
- Subjects were trained in the high condition only
- Needed d' > 1 in one 40-trial training block to pass

4. RESULTS (cont.)

Localization Conditions

Each color represents a subject. The circles represent the mean for each condition along with 95% confidence intervals. The average performance in the Low+High condition is significantly worse than the High and Stream conditions (p < 0.01 for each, single-tailed t-test)

0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0

Percent Correct

High Low + High Stream Noise

4. RESULTS

Identification Conditions

Each color represents a subject. The circles represent the mean for each condition along with 95% confidence intervals. The average performance in the Low+High condition is significantly worse than that in the High and Stream conditions. Also, the noise control condition performance is significantly better than that in the Low, High and Stream conditions. The average performance in the Low+High condition is significantly better than that in the Low, High and Stream conditions (p < 0.01 for each, single-tailed t-test).

5. CONCLUSIONS

- Automatic grouping processes operate on speech objects, influencing the perception of different attributes, such as identity and location.
- The effects of grouping can either degrade or enhance performance, depending on the task.

6. ACKNOWLEDGMENTS

This work was supported by NIDCD grant R01 DC00947.

7. REFERENCES

