Coherence masking protection for mid-frequency formants by adults and children

Eric Tarr and Susan Nittrouer
Department of Otolaryngology, The Ohio State University, 915 Olentangy River Road, Suite 4000, Columbus, Ohio 43212
tarr.18@osu.edu, nittrouer.1@osu.edu

Abstract: Coherence masking protection (CMP) refers to the phenomenon in which a target formant is labeled at lower signal-to-noise levels when presented with a stable cosignal consisting of two other formants than when presented alone. This effect has been reported primarily for adults with first-formant (F1) targets and F2/F3 cosignals, but has also been found for children, in fact in greater magnitude. In this experiment, F2 was the target and F1/F3 was the cosignal. Results showed similar effects for each age group as had been found for F1 targets. Implications for auditory prostheses for listeners with hearing loss are discussed.

PACS numbers: 43.66.Ba, 43.71.An, 43.71.Ft [QJF]
Date Received: July 12, 2011 **Date Accepted:** August 18, 2011

1. Introduction

Perceptual mechanisms responsible for auditory processing across broad frequency regions can help with segregating signal components from noise, and with grouping signal components into coherent auditory objects (Bregman, 1990; Darwin and Carlyon, 1995; Yost and Sheft, 1994). In a well-known demonstration of these effects, a signal is recognized in background noise at lower levels when noise outside the critical band of that signal is presented with the same temporal envelope as noise within the critical band (e.g., Hall and Grose, 1990; Hall, *et al.*, 1984). This phenomenon, termed comodulation masking release (CMR), illustrates how a mechanism that facilitates the perceptual grouping of masker components affects detection of the signal. One advantage of across-frequency auditory processing is that signals can be detected at lower levels when masker components cohere perceptually.

Gordon (1997) modified procedures used to study CMR in ways necessary to examine how perceptual grouping mechanisms might help listeners recognize speech signals in noise. In this paradigm, listeners are asked to label an isolated first formant (F1) as belonging to either a high or low front vowel (/ɪ/ or /ɛ/) in the presence of noise filtered to cover the critical band of that F1. Thresholds for accurate *labeling* are measured, which differs from procedures used to demonstrate CMR where thresholds for *detection* are measured. In a second condition, speech components outside that critical band—an F2/F3 complex—are presented synchronously with F1. Although this "cosignal" is outside the critical band, remains stable across trials, and is low in amplitude, listeners accurately label vowel stimuli at poorer signal-to-noise ratios when it is present. This phenomenon is termed coherence masking protection (CMP) and is another demonstration of the benefits provided by across-frequency perceptual mechanisms: In this phenomenon, protection against masking is obtained when components of the signal cohere. Thus, CMP differs from CMR in that the latter depends on facilitating coherence in the masker. But regardless of whether it is masker or signal components that are being coaxed into cohering across a broad frequency array, release, or protection, from masking is obtained.

The procedures and outcomes of experiments on CMP could provide lessons about how to help listeners with hearing loss. Of most relevance is the fact that CMP

has clearly and most often been demonstrated for speech signals. Improving speech recognition in noise for listeners with hearing loss has been one of the toughest problems to solve, more difficult than improving speech recognition in quiet for these listeners. Further, in CMP the additional elements that need to cohere to the signal component that determines phonetic identity can be presented at levels barely above threshold and the effect holds. This means that CMP could potentially aid recognition for listeners with hearing loss where audibility is the chief obstacle to treatment. In a nutshell, the very demonstration of CMP for speech stimuli suggests that promoting audibility across a broad frequency range might be beneficial to speech recognition in noise for listeners with hearing loss. Even at frequencies where thresholds are quite poor there could be an advantage to trying to amplify the signal just enough so that it remains slightly above aided thresholds, rather than transposing or compressing the frequency range of amplified signals, which is one clinical option being recommended at present (e.g., Bohnert *et al.*, 2010).

But at least one important question remains unanswered concerning CMP. To date, demonstrations of CMP for speech signals have only been obtained using synthetic versions of front vowels differing in vowel height. Work by both Gordon (1997, 2000) and Nittrouer and Tarr (2011) used the vowels /ɪ/ and /ɛ/ because using front vowels as targets makes it easy to ensure that the F2/F3 cosignal is well outside of the critical band of those F1 frequencies. Making vowel height the attribute to be labeled also means that only F1 frequency needs to be manipulated; the cosignal can remain stable across stimuli without influencing vowel recognition. Nonetheless, this choice of stimuli constrains the relevance of the phenomenon for understanding speech perception by listeners with hearing loss. Vowel height is less likely to pose problems for listeners with hearing loss than is vowel place (Boothroyd, 1984), a fact that arises precisely because vowel height is signaled in the acoustic structure of vowels by the frequency of F1. This formant is low frequency, and hearing loss is typically poorer in the higher frequencies than in the lower frequencies when the loss is anything other than flat. And even when hearing loss is so severe in the low frequencies that audibility of F1 is affected, jaw height is visible on the face, making vowel height recognizable that way. Thus, if coherence of signal components protects against masking for only low-frequency speech components, the effect would be of little utility to the treatment of hearing loss. The purpose of the study reported here was to see if CMP is found when the target is higher in frequency, signaling vowel place.

In the current study, F2 served as the target. Vowel stimuli were selected to be midway in height on the vowel quadrilateral, both to avoid needing to invoke the effects of lip rounding, which is present for high, back vowels in English, and to avoid having F1 be high, as happens for vowels that are low (or open). The latter factor would make it hard to separate F1 and F2 sufficiently. A major goal of the study was to see if the magnitude of the CMP effect would be similar to what has been reported when F1 is the target formant. In several studies, Gordon (1997, 2000) observed CMP effects for adults of roughly 3.2 dB, meaning that the labeling threshold was that much lower when the F1 targets were presented coincidentally with the F2/F3 cosignal. Nittrouer and Tarr (2011) replicated this finding, reporting an effect of 3.3 dB for adults listening to synthetic speech with harmonically related targets and cosignals. An effect of this size might seem small, until it is viewed as an effective improvement in signal-to-noise ratio (SNR) of 3.3 dB. Studies of word recognition in noise have demonstrated that an improvement of 1 dB in SNR increases recognition of individual words by about 6.6% (e.g., Boothroyd and Nittrouer, 1988). Viewed from this perspective, 3.3 dB of masking protection could offer a real advantage to listeners with hearing loss.

Finally, Nittrouer and Tarr (2011) measured CMP for children of two ages: 8 and 5 years. The magnitudes of CMP effects were even greater for these listeners than they were for adults: 6.2 dB for 8-year-olds and 9.2 dB for 5-year-olds. The present study sought to examine whether this age-related trend would be replicated when the

target formant was F2, rather than F1. If so, it would mean that having broad frequency components available to cohere into unitary percepts is especially important for children. Conversely, it would suggest that any sort of frequency compression that is done on signals for presentation through hearing aids might be particularly detrimental to children. This would be important information to have when it comes to the clinical treatment of hearing loss.

In summary, the current study sought to extend the work of Nittrouer and Tarr (2011) by examining whether or not CMP would be obtained for speech signals when F2 is the target formant upon which vowel recognition depends. Adults and children participated in order to see if children demonstrated the enhanced effects in this configuration that they had shown when F1 was the target.

2. Method

2.1 Listeners

Sixty-eight listeners participated in this experiment. All subjects (or in the case of children, their parents on their behalf) reported English as their native language and stated they had no history of speech, language, or hearing problems. A hearing screening was performed using the frequencies of 0.5, 1.0, 2.0, 4.0, and 6.0 kHz presented at 25 dB hearing level to each ear separately as a minimum hearing requirement. Children were included only if they had fewer than five episodes of otitis media before the age of 3 years. The distribution of participants tested by age was: 20 adults between 18 and 37 years; 23 children between 8 years, 0 months and 8 years, 11 months; and 25 children between 5 years, 0 months and 5 years, 11 months.

2.2 Equipment and materials

Hearing screenings were performed with a Welch Allen TM-262 audiometer and TDH-39 earphones. The experiment was done using AKG-K141 headphones, a Samson Q5 amplifier, and a Soundblaster digital-to-analog converter. The computer that controlled stimulus presentation was outside the sound-isolation booth used for testing.

Listeners pointed to one of two pictures on cardboard (6 in. $\times$ 6 in.), each of which corresponded to a response choice. One picture was of a closed flower (*bud*), and the other was of a bed (*bed*).

2.3 Stimuli

Two synthetic vowel stimuli were created with a software synthesizer, Sensimetrics "SenSyn," using a 10 kHz sampling rate and 16 bit digitization. Both vowels were 60 ms long, including 5 ms on and off ramps, consisted of three formants, and had a constant fundamental frequency (f0) of 125 Hz. These vowels were modeled after /ʊ/ and /ɛ/. The first and third formants (F1 and F3) were 600 and 2400 Hz for both vowels. The second formant (F2) was 1125 Hz for /ʊ/ and 1875 Hz for /ɛ/. Formant bandwidths (at 3 dB below peak amplitude) were 50 Hz for F1, 110 Hz for F2, and 170 Hz for F3.

These vowel tokens were subsequently manipulated to create two versions of each vowel: F2-only and full-formant stimuli. To create the F2-only stimuli, each vowel was band-pass filtered to preserve only the midfrequency regions where F2 is located. Specifically, the three-formant stimuli were band-pass filtered using a digital Hamming filter with 50 dB attenuation for frequencies below 800 Hz, a transition band to 900 Hz, a pass band between 900 and 2050 Hz, a transition band to 2 150 Hz, and 50 dB attenuation for frequencies above 2150 Hz. The /ɛ/ vowel was used to recover the F1 and F3 portions, separately, for use in creating the full-formant stimuli. The F1 portion was recovered by low-pass filtering the three-formant /ɛ/ stimulus below 800 Hz, with a transition band to 900 Hz, and 50 dB attenuation above 900 Hz. The F3 portion was recovered by high-pass filtering the signal above 2150 Hz, with a transition band down to 2050 Hz, and 50 dB attenuation below 2050 Hz. Then the two full-formant stimuli were created by combining the F1 and F3 portions with the

band-pass F2-only stimuli using synchronous onsets and offsets for all three parts. Creating these stimuli in this way allowed precise control over the amplitude relations of F2 and the other two formants. Full-formant stimuli were created with the F1 portion 6 dB lower than F2, and the F3 portion 12 dB lower than F2. Pilot work by us with 24 adults showed no differences in outcomes from those reported here when the amplitude of the F1 cosignal varied between 0 and 12 dB below F2 in 6 dB steps. F3 was maintained at a level 12 dB below that of F2 through all pilot testing. Figure 1 shows smoothed spectra of the full-formant stimuli.

Synthetic versions of the words *bud* and *bed* used in training were created from the full-formant stimuli by appending formant trajectories at the beginnings and ends. The word *bud* was created from /ʊ/ by appending 40 ms transitions at the beginning with starting frequencies of 300, 800, and 1900 Hz for F1, F2, and F3, respectively. The same was done to create *bed* from /ɛ/, but F2 started at 1500 Hz. Steady-state syllable portions were 100 ms. At the end of the word *bud*, 40 ms transitions were appended, with ending frequencies of 300, 1425, and 2600 Hz for F1, F2, and F3, respectively. The same was done for *bed*, but F2 ended at 2175 Hz.

A pseudo-random-number generator in MATLAB was used to generate flat-spectrum white noise. The masking noise was 600 ms long, and was band-pass filtered between 900 and 2050 Hz using the same filter specifications as those used to create the F2-only stimuli.

2.4 Procedures

The order of presentation of the two kinds of stimuli was chosen such that half the participants started with F2-only stimuli followed by full-formant stimuli and the other half had the opposite order of presentation.

Training: All subjects started with the same two-step training process. The first step consisted of the word stimuli. The synthetic words were presented one at a time over headphones at 74 dB sound pressure level (SPL) in random order without noise. The listener both pointed to the correct picture and said the correct word because having both kinds of responses served as a check on whether listeners (especially children) understood which label went with which picture. Feedback was provided. Fifty words were presented. A correct response was recorded when the listener pointed to the appropriate picture and said the correct word. When a mismatch in pointing and naming occurred, which was rare, listeners were reminded that the picture they point to must match the word they say.

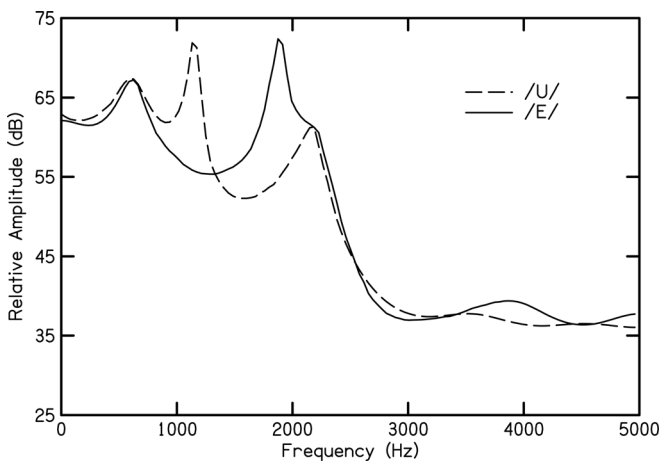

Fig. 1. Spectra of full-formant stimuli.

After synthetic word training, all listeners performed full-formant vowel training, which consisted of 50 presentations of the 60 ms synthetic vowels at 74 dB SPL in random order without noise. Listeners continued to label each stimulus by pointing and saying the entire word that the vowel was part of. Feedback continued during full-formant vowel training.

If the listener was to start with the full-formant stimuli during testing, they would next do the pretest for this condition immediately following this training. If the listener was to start with the F2-only condition, they would next complete training with 50 tokens of F2-only stimuli in quiet with feedback and perform the pretest with those stimuli. Training and the pretest with the full-formant stimuli would be provided prior to testing with those stimuli.

Pretest trials: After training in each condition, but prior to testing, listeners were required to label nine out of ten correct stimuli (without noise) of the kind to be used in testing. A maximum of 50 stimuli were presented without feedback. As soon as the listener responded correctly to nine out of ten consecutive presentations, the pretest stopped. If 50 stimuli were presented without the listener ever responding correctly to nine out of ten consecutive presentations, that listener was not tested in that condition and all data from that listener were eliminated from the analysis.

Adaptive testing: For each stimulus condition, an adaptive procedure (Levitt, 1971) was used to find the SNR at which each listener could provide the correct vowel label 79.4% of the time. The noise was held constant at 62 dB SPL throughout testing, and the level of the signal varied. The initial signal level was 74 dB SPL. After three consecutive correct responses, the level of the signal decreased by 8 dB. Signal level continued to decrease by 8 dB after three consecutive correct responses until the listener made one labeling error, which marked the end of the first run. The level of the signal then increased by 8 dB each time the listener gave a wrong response until the listener responded correctly to three consecutive stimuli. At that time the change in signal level began to decrease again. The signal level changed by 8 dB after the first two runs, at the reversals and then the signal level changed by 4 dB at the next two reversals. For the final 12 reversals the signal level changed by 2 dB. The signal levels of the last eight reversals were averaged and used as the labeling threshold. Feedback was not provided during adaptive testing. The order of stimulus presentation was randomized by the software.

Post-test trials: After testing was completed, listeners heard ten stimuli without noise and without feedback. Failure to correctly label nine out of ten stimuli resulted in a listener's data being excluded from analysis. Listeners had to meet the pre- and post-test inclusionary criteria for all conditions in order for their data to be included in any condition. After completing the post-test for the first condition (F2-only or full-formant), listeners began training for the other condition, starting with 50 tokens with feedback.

3. Results

Three 8-year-olds (13%), and five 5-year-olds (20%) failed to meet the pre- or post-test criterion for one or both conditions. All but one of those 5-year-olds failed to meet the criterion for the F2-only condition, but passed with the full-formant stimuli. The other 5-year-old failed with both the F2-only and full-formant stimuli during the post-test. Two of the three 8-year-olds who were excluded failed the post-test for both conditions. The other 8-year-old failed the post-test for just the F2-only condition. Results from 20 listeners of each age were included in data analysis.

Table 1 shows mean labeling thresholds and standard deviations by age group for each of the conditions. The difference between condition thresholds, or magnitude of the CMP effect, increased with decreasing age. Means for CMP effects computed on individual thresholds were 3.7 (3.0) for adults, 5.6 (3.7) for 8-year-olds, and 7.1 (3.1) for 5-year-olds.

A two-way analysis of variance was performed on the thresholds shown in Table 1, with age as a between-subjects factor and number of formants (one or three)

as a within-subjects factor. The main effects of age, $F(2,57) = 21.25$, $p < 0.001$, and number of formants, $F(1,57) = 165.09$, $p < 0.001$ were both significant. The Age × Formants interaction was also significant, $F(2,57) = 5.03$, $p = 0.001$, reflecting the fact that the magnitude of the CMP effect increased with decreasing age.

Labeling thresholds and CMP effect sizes were similar to those reported by Nittrouer and Tarr (2011) for synthetic speech. In that earlier study, mean thresholds for adults were 61.2 and 57.9 dB SPL for F1-only and full-formant stimuli, respectively, with a mean CMP effect of 3.3 dB. Mean thresholds for 8-year-olds were 65.0 and 58.8 dB SPL for F1-only and full-formant stimuli, respectively, with a mean CMP effect of 6.2 dB. Mean thresholds for 5-year-olds were 70.2 and 61.1 dB SPL for F1-only and full-formant stimuli, respectively, with a mean CMP effect of 9.1 dB. Comparing those earlier outcomes to findings from the current study (Table 1) makes it clear that adults and 8-year-olds performed almost identically across the two studies. Thresholds were actually a little lower in the current study for 5-year-olds than in the earlier study, but disproportionately so for the F2-only condition. This latter fact meant that the mean CMP effect was reduced slightly from what was reported by Nittrouer and Tarr (2011).

4. Discussion

The current study was undertaken to see if the across-frequency perceptual mechanisms that operate so effectively to lower thresholds in noise for low-frequency speech targets would operate equally as well when the target is midfrequency and the cosignal straddles that target. In earlier investigations, the target was always F1 and the cosignal to be integrated consisted of higher formants. Outcomes of those experiments showed that listeners group formants into coherent auditory objects, a perceptual strategy that serves to lower thresholds for accurate phonetic labeling when the target formant is presented in noise. The magnitude of masking protection observed with this paradigm was greater for children than for adults.

Results of the current study demonstrated that equivalent effects are observed when the target formant is F2, and the cosignal is F1 and F3. Again, the magnitude of the effect was greater for children than for adults.

These outcomes have implications for the design and fitting of auditory prostheses in listeners with hearing loss. For decades, traditional approaches to understanding speech perception have focused on the idea that separate sensory channels tuned to narrow frequency regions allow listeners to recover discrete bits of information, which are used for phonetic processing. In these experiments on CMP we see that the auditory system operates with mechanisms encompassing broader frequency ranges. Being capable of forging auditory objects that span wide regions of the spectrum supports better processing of phonetically relevant signals. This is true regardless of where the most informative signal component lies on the spectrum. When this perspective of perceptual processing is kept in mind, it suggests that as much of the spectrum as possible should be preserved and amplified for listeners with hearing loss. The study reported here did not examine speech perception by listeners with hearing loss directly, but nonetheless suggests that efforts to compress the speech spectrum for those listeners should proceed cautiously. Prosthesis developers and clinicians should remain mindful

Table 1. Means (*M*) [and standard deviations (SD)] of labeling thresholds.

Age		F2 only		Full formant	
	n	*M*	SD	*M*	SD
Adults	20	61.0	1.8	57.3	2.0
8-year-olds	20	64.7	3.8	59.1	1.4
5-year-olds	20	67.4	3.7	60.3	3.1

of the benefits of these across-frequency mechanisms. This concern applies especially to children, who benefit even more than adults from having broad spectral components of the signal accessible. Although the goal of frequency compression is largely to make fricative noises available to listeners—a particularly important purpose for children who are still acquiring language—that goal needs to be balanced with the benefits that clearly derive from having broad frequency signals. The current study did not examine CMP for nonspeech signals, but future work should explore the applicability of the phenomenon to these sorts of signals. It may very well be that maintaining broad spectral arrays is as important to other kinds of signals, such as music, as it is for speech.

Acknowledgments

This work was supported by a grant from the National Institutes of Health, National Institute on Deafness and Other Communication Disorders, R01 DC000633, to S.N.

References and links

Bohnert, A., Nyffeler, M., and Keilmann, A. (**2010**). "Advantages of a non-linear frequency compression algorithm in noise." Arch. Otorhinolaryngol. **267**, 1045–1053.

Boothroyd, A. (**1984**). "Auditory perception of speech contrasts by subjects with sensorineural hearing loss." J. Speech Hear. Res. **27**, 128–134.

Boothroyd, A., and Nittrouer, S. (**1988**). "Mathematical treatment of context effects in phoneme and word recognition." J. Acoust. Soc. Am. **84**, 101–114.

Bregman, A. S. (**1990**). *Auditory Scene Analysis* (The MIT Press, Cambridge, MA).

Darwin, C. J., and Carlyon, R. P. (**1995**). "Auditory grouping," in *The Handbook of Perception and Cognition*, Vol. 6 Hearing, edited by B. C. J. Moore (Academic, San Diego, CA), pp. 387–424.

Gordon, P. C. (**1997**). "Coherence masking protection in speech sounds: The role of formant synchrony." Percept. Psychophys. **59**, 232–242.

Gordon, P. C. (**2000**). "Masking protection in the perception of auditory objects." Speech Commun. **30**, 197–206.

Hall, J. W., and Grose, J. H. (**1990**). "Comodulation masking release and auditory grouping." J. Acoust. Soc. Am. **88**, 119–125.

Hall, J. W., Haggard, M. P., and Fernandes, M. A. (**1984**). "Detection in noise by spectro-temporal pattern analysis." J. Acoust. Soc. of Am. **76**, 50–56.

Levitt, H. (**1971**). "Transformed up-down methods in psychoacoustics," J. Acoust. Soc. Am. **49**, 467–477.

Nittrouer, S., and Tarr, E. (**2011**). "Coherence masking protection for speech signals in children and adults," Atten. Percept. Psychophys. doi:10.3758/s13414-011-0210-y.

Yost, W. A., and Sheft, S. (**1994**). "Modulation detection interference: across-frequency processing and auditory grouping." Hear Res. **79**, 48–58.