Cross-categorical discrimination of simple speech and music sounds based on timbral fidelity in musically experienced and naïve listeners

Ryan Anderson[1†], Alyxandria Sundheimer[1] and William P. Shofner[1]
[1]Indiana University, Bloomington Indiana, United States
[†]Corresponding author: anderyan@iu.edu

Introduction

Psychoacoustic approaches to complex sound perception suggest that there are differences in how normal hearing humans use spectral information in speech and music signals. Studies applying these approaches conclude that music perception requires greater spectral resolution than speech perception (Shannon, 2005). Intelligibility of noise-vocoded speech is high with as few as 4 vocoder channels (Smith et. al, 2002). Conversely, music perception studies with noise vocoded signals suggest that upwards of 32 channels are necessary for recognition (Mehta and Oxenham, 2014). Analyzing these results together, it seems that music perception is more susceptible to spectral degradation than speech perception. That is, a high level of speech perception performance can be achieved with fewer noise-vocoded channels than required for music perception performance. Such conclusions support arguments for specialized cognitive processes in which the brain uses acoustic information differently depending on the type of sound its processing. This logic propels popular theories regarding auditory processing specialization at various cognitive levels (Liberman, 1984; Zatorre, 2002). However, conclusions from these metadata are problematic given that they aggregate results from several different studies using different methodologies and therefore different cues. In particular, music perception as represented by melody recognition relies on changes in pitch information and the structure of harmonic information across the duration of the stimulus. Conversely, word identification tasks use envelope cues generated by the spectral information in consonants and formant structure of vowels. Given that noise vocoding is used to introduce spectral content manipulations, it is important to note that these stimuli are influenced differently.

Differences in speech and music perception are also prevalent in studies regarding subjects' musical experience. These studies generally demonstrate that musical experience correlates with better speech perception in degraded or challenging conditions (Parbery-Clark et al. 2012) as well as frequency and pitch discrimination (Tervaniemi et. al, 2005) compared to musically inexperienced peers. Based on these differences, the level of musical experience in participants should be considered when exploring differences in categorical sound perception.

Vocoding affects speech and music differently depending on their task context. Furthermore, it is unclear as to whether specific or general mechanisms are driving decisions due to different task demands. To eliminate task differences, the experiment at hand evaluates the perception of speech and music sounds using a single task in which the acoustic cues are equivalent in all conditions. By using natural representations of spoken vowels and music notes, respective spectral structures serve as the primary differentiating acoustic feature across stimuli. Spectral profiles of harmonic structure in musical instruments and formant distribution in vowels provide a common dimension of timbre between speech and music. Therefore, a behavioral task in which the participants make assessments of timbral differences across categories of speech and music in natural and vocoded conditions is used to determine processing differences as dependent on the spectral quality of the signal. The present study expands on preliminary data using a single discrimination paradigm to compare speech and music perception based on similar perceptual dimensions, namely timbre.

Method

28 subjects with normal hearing thresholds (< 20 dB HL across audiological test frequencies) completed experiment 1. Of these participants, 17 were naïve listeners and 11 were musically experienced. Musically experienced participants were considered as those who reported 3 or more years of practiced musical

experience. Regardless of classification, all participants completed all trials of each experiment. 6 subjects completed experiment 2; three of which completed experiment 1.

Stimuli consisted of naturally spoken vowels and notes played on musical instruments as well as 32-, 8-, and 4- channel noise-vocoded versions. Music note samples consisted of bassoon, cello, clarinet, trombone, trumpet, and viola playing either G3 (196 Hz), or B2 (123 Hz). Vowels */a/, /ae/, /i/, /ou/, /u/* were recorded from male and female speakers. All stimuli were equalized in RMS amplitude and presented for 500 ms at 73 dB SPL. Fundamental frequencies for music notes and vowels were closely paired within trials to remove potential pitch cues. Listeners discriminated either instruments from vowels or vowels from instruments via button release in a go/no-go task. Before the participants initiated a trial, a 500 ms standard token was repeatedly presented with a 500 ms interstimulus interval. Participants prompted a trial by pressing and holding down a button. While holding down the button, an 1850 ms response window was placed after a randomly selected hold time (1150 – 8150 ms). During the response window, two 500 ms sounds were presented with a 500 ms interstimulus interval. Participants were instructed to only release the button if the sound source in the response window was from the other category than the category containing the standard token (i.e. if the standard token is a vowel, only release the button if a music note plays). A button release to a within category change indicates a false alarm. The discrimination paradigm offered insight as to how timbral fidelity influenced perception *between* stimulus sound categories. Reaction times and accuracy were measured and organized as a function of signal degradation level to consider potential differences in how normal hearing listeners utilize spectral information. To clarify the cause of changes in task performance, a second experiment used the same procedure as experiment one but tasked the participants with responding to any perceived change from the standard stimulus, regardless of category. If participants are unable to accurately discriminate targets from other stimuli solely due to the general spectral quality of the sound, one would expect for performance in both experiments to be similar. Trials consisted of just 8- and 4-channel vocoded conditions, as these are the most challenging and therefore had the largest chance to influence categorization ability.

Results

Experiment 1

Grand average reaction times in response to 4 channel, 32 channel, and natural stimuli showed no difference beyond error in both conditions (figure 1). Like in preliminary data, there is a significant effect of spectral quality of the signal on accuracy measures with decreasing accuracy as the number of vocoding channels decreases in conditions where discrimination occurs across categories (figure 2). This is

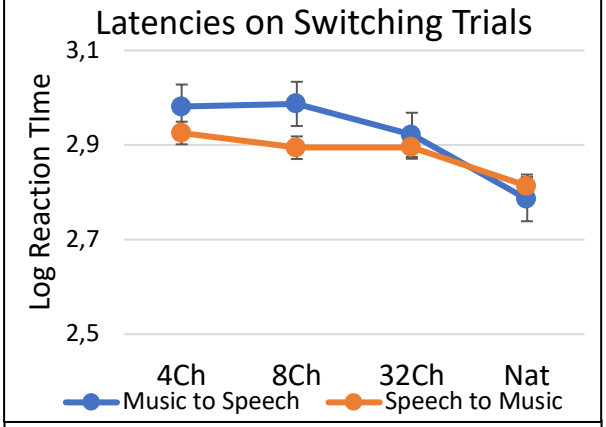

Fig 1. Grand average log transformed reaction times for both conditions as a function of number of vocoding channels.

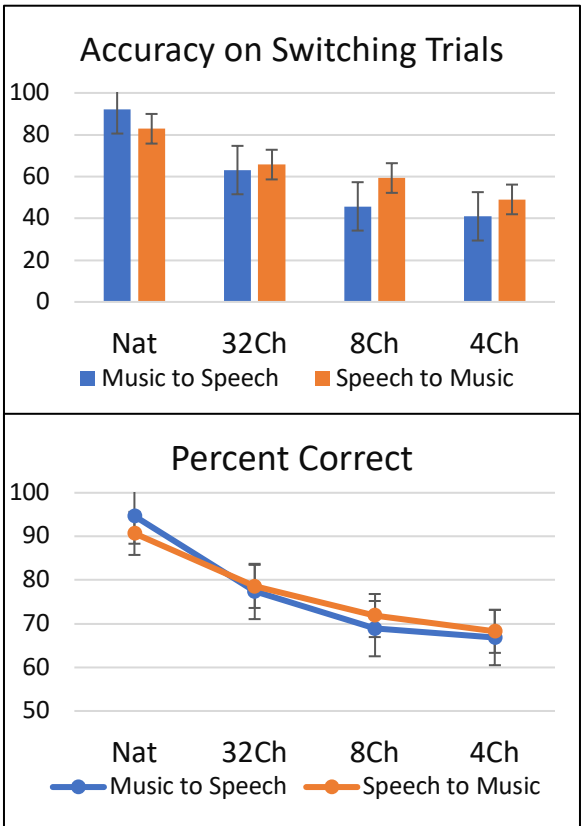

Fig 2. Grand average accuracy on cross-category discrimination trials as a function of vocoding channels. Hit rates in both conditions (top panel) and percent correct (bottom panel) compare performance in both category conditions.

expected, as fewer channels generate a more ambiguous signal. Within vocoding conditions however, there are no significant differences in accuracy measures when discriminating across categories. Using discrimination trials across all listeners as an indication, there are no substantial differences beyond error in discrimination accuracy between perception of basic speech and music sounds, regardless of spectral richness.

Preliminary data suggest that musically trained listeners are better than naïve listeners at discriminating vowels from musical instruments (Anderson et. al, 2019). After including more musically trained subjects in analysis, this asymmetry was still observed, but to a smaller effect compared to the previous study. Two-factor ANOVA on arcsine transformed percent correct with Bonferroni corrections showed a significant effect of group in the music-to-speech condition, $F(1,104) = 7.63$, $p < .01$; $\eta^2 = .07$), while there was no significant effect of group in the speech-to-music condition (figure 3).

Experiment 2

In the general task, participants demonstrated a clear ability to discriminate between the stimuli in both vocoding representations. Percent correct scores were near ceiling for 8 channel noise vocoded stimuli (99.68%), and at ceiling for 4 channel stimuli. These data ensure that participants are indeed able to differentiate the stimuli from experiment 1 even with the largest amount of spectral degradation in the stimuli set. All effects observed in experiment 1 should therefore be attributed to differences in category discrimination ability.

Discussion

Using a single categorical discrimination task in which timbre was the primary decision criteria allows for a more justifiable basis of comparison between the role of spectral processing in speech and music sounds. This type of design ensures that the acoustic cues as well as the task demand are equivalent. Given the balancing of acoustic cues in this design, asymmetrical performance potentially reflects differences in perceptual modes that depend on the type of stimuli being processed. Lower vocoding channel stimuli are of interest, as these localize the previously reported split in performance between music and speech processing. Slight differences in reaction times for spectrally ambiguous stimuli (8-channels) suggest a potential deviation in speech and music processing, but accuracy measures do not currently support this possibility, as there were no substantial differences in average responses in any vocoding conditions. Similarities across listeners in accuracy and reaction time measures when discriminating between sound categories indicate that vowel and musical instrument identification do not involve mode-specific mechanisms. When accounting for musical experience of the listeners, there is a persistent trend in which musically trained listeners demonstrate higher percent correct responses than naïve listeners at

discriminating vowels from musical instruments. This is a curious finding as one would expect the musically trained group to exhibit similar reaction times and percent correct measures in trials where the target switches across conditions regardless of discrimination direction. Reaction times should be significantly lower, and percent correct significantly higher compared to naïve listening peers if this were the case. Given the decrease in effect size from the preliminary study and the current study as the number of subjects increased, it is not unlikely that this asymmetrical trend might be eliminated with larger, matched group sizes. Future studies using more stimuli in each category and larger matches samples may provide greater validity and statistical power to investigate this possibility.

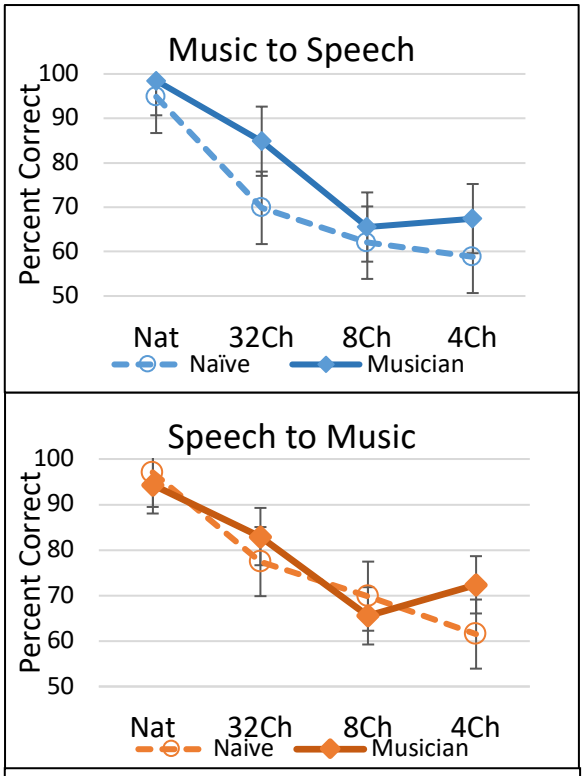

Fig 3. Group level performance as indicated by percent correct in music to speech (top) and speech to music (bottom) conditions.

References

Anderson, R., Sundheimer, A., & Shofner, W. (2019). Ability of normal hearing listeners to recognize vowels and musical instruments under spectrally-degraded conditions. *The Journal of the Acoustical Society of America*, *145*(3), 1720-1720.

Liberman, A. M. (1984). On finding that speech is special. In *Handbook of Cognitive Neuroscience* (pp. 169-197). Springer, Boston, MA.

Mehta, A. H. & Oxenham, A. J. (2017). Vocoder simulations explain complex pitch perception limitations experienced by cochlear implant users. *Journal of the Association for Research in Otolaryngology, 18,* 789 – 802.

Parbery-Clark, A., Tierniey, A., Strait, D.L. & Kraus, N. (2012). Musicians have fine-tuned neural distinction of speech syllables. *Neuroscience, 219,* 111-119.

Shannon, R. V. (2005). Speech and music have different requirements for spectral resolution. *International Review of Neurobiology, 70,* 121-134.

Smith, Z. M., Delgutte, B., & Oxenham, A. J. (2002). Chimaeric sounds reveal dichotomies in auditory perception. *Nature*, *416*(6876), 87-90.

Tervaniemi, M., Just, V., Koelsch, S., Widmann, A., & Schröger, E. (2005). Pitch discrimination accuracy in musicians vs nonmusicians: an event-related potential and behavioral study. *Experimental brain research*, *161*(1), 1-10.

Zatorre, R. J., Belin, P., & Penhune, V. B. (2002). Structure and function of auditory cortex: music and speech. *Trends in cognitive sciences*, *6*(1), 37-46.