5-2016

Effects of talker variability on spectral contrast effects.

Asim Mohiuddin

Effects of Talker Variability on Spectral Contrast Effects

By

Asim Mohiuddin

Submitted in partial fulfillment of the requirements
for Graduation *summa cum laude*
and
for Graduation with Honors from the Department of Psychology

University of Louisville

March, 2016

Abstract

Spectral contrast effects are context-dependent effects that influence the way we perceive certain sounds. Evidence of these effects can be seen in experiments where a precursor sound (e.g. a sentence) is followed by a target vowel sound (like /ɪ/ as in "bit" or /ɛ/ as in "bet"). If the precursor's frequency was emphasized in areas more consistent with the frequency of /ɛ/, listeners tend to perceive the target sound to be the opposite i.e. /ɪ/. A recent study shows using sentence precursors from 200 different talkers diminishes these effects questioning previous claims that talker variability has no influence on spectral contrast effects (Assgari & Stilp, 2015; Lain, Liu, Lotto, & Holt 2012). This study investigated the influence of one, four, eight, and sixteen talkers using conversational speech. Sentences were filtered with +5 dB filter gain to emphasize frequency regions consistent with either /ɛ/ or /ɪ/. One-talker, four-talker, and eight-talker conditions all produced contrast effects while the sixteen-talker condition failed to produce an effect. Results suggest talker variability has a greater influence on spectral contrast effects than previously thought.

Effects of Talker Variability on Spectral Contrast Effects

Speech perception is a complex task that involves deciphering words from a broad range of sounds. Whether we are listening to a high-pitched voice in an echoing lecture hall or a man on a noisy street, we are able to understand speech with relative ease. Not only do we extract speech from this general sound signal, but we also categorize the different speech sounds into vowels so we may understand the speech. This poses a question of what mechanisms we use in categorizing vowels within speech that lead to our perception. Is our auditory system equipped with the ability to categorize vowel sounds based on their absolute frequencies or is speech perceived based on relative frequencies of a talker's sounds? The latter theory was proposed by Joos (1948) and was later used to explain experiments done by Ladefoged & Broadbent (1957) who showed that listeners identified the same word differently after listening to a sentence whose formant frequencies were modified. The current study seeks to explore the effects described by Ladefoged & Broadbent (1957), now known as spectral contrast effects, in relation to whether the number of talkers has an influence on our experience of these contrast effects.

In 1957, Ladefoged and Broadbent published a seminal paper that showed categorization of vowels can be influenced by the context in which the vowel sounds are placed. They performed an experiment where they created variations of the sentence "Please say what this word is," manipulating its formant frequencies. The formant frequency is the frequency at which a sound reaches a relative maximum or a spectral peak and allows us to differentiate between vowels. In their experiment, they modified the sentence to shift the formant frequency towards the low region (towards a vowel sound like /ɪ/ as in "bit") or the high region (towards a vowel sound like /ɛ/ as in "bet") of the first formant (F_1). They then synthesized target words like "bit" and "bet". Participants were asked to distinguish these target words after listening to the

manipulated sentences. Following an unmodified sentence, almost all participants judged the word in question as intended. When participants heard the sentence modified to emphasize the frequency of /ɛ/ (high F_1), almost all participants judged the test word to be "bit" regardless of whether "bit" or "bet" was played. Similarly, when "bit" or "bet" was preceded by a sentence with /ɪ/ peaks emphasized (low F_1), participants almost always judged the test word to be "bet" (Ladefoged & Broadbent, 1957). The tendency for people to judge the test word as opposite of what was emphasized in the preceding sentence is what is now known as the spectral contrast effect (Stilp, Anderson, & Winn, 2015).

The spectral contrast effect has been interpreted and explained in many ways. One explanation that seeks to reconcile different reports of this effect explains it in terms of sensory adaptation (Stilp et al., 2015). Our sensory system tends to become less sensitive to stable stimuli in a given context (Barlow, 1961; Stilp et al., 2015). This allows us to be able to quickly perceive any changes in stimuli. In other words, we tend to perceptually magnify the difference between stable properties of a stimulus and any changes in the stimulus (Stilp et al., 2015). In context of Ladefoged & Broadbent (1957), this means that when participants adapted to the stable spectral peaks emphasizing a certain vowel in a sentence, they magnified the difference between the test word and the preceding sentence. This shifted their perception to the opposite direction of what was emphasized in the preceding sentence.

There have been many studies showing this effect is very robust and can be produced for a wide range of sounds. Watkins (1991) found that contrast effects hold for manipulations like presenting the test vowels and precursor sentence to different ears, using male and female sentences followed by male test vowels, and when sentences were reversed to produce unnatural sounds. This effect can also be observed in English-Spanish bilingual, Spanish-, English-, or

Dutch-speaking listeners using different native and non-native languages (Sjerps & Smiljanić, 2013). There have also been a number of studies that show spectral contrast effects extend beyond speech sounds. For example, contrast effects hold for musical instruments where instead of test words, participants heard a tenor saxophone and French horn (Stilp, Alexander, Kiefte, & Kluender, 2010). Even simple compilations of sine tones (or "tone histories") that vary in duration and number of tones can be used to produce the effect (Holt, 2005; Holt, 2006).

Traditionally, sentences or precursor sounds that are used in spectral contrast effect experiments are modified by introducing relatively large peaks [e.g. +30 decibel (dB) peaks] in key frequency regions (see Stilp et al., 2015). Such large peaks may not be common in natural settings (Assgari & Stilp, 2015). In experiments done by Stilp et al. (2015), a significant contrast effect was observed with as small as +5 dB filter gain. However, using such a small filter gain resulted in a significantly smaller effect when compared to the effect observed with +20 dB filter gain (Stilp et al., 2015).

A major question surrounding spectral contrast effects is whether it is purely an acoustic phenomenon or it can be subject to higher-level influences like talker information. In the light of the above studies, it is clear that spectral contrast effects can be produced in a variety of contexts. Laing et al. (2012) observed contrast effects when they adjusted formant frequencies to create sentence precursors that perceptually appear to be produced by different talkers. The authors claimed that talker information plays no role in spectral contrast effects (Laing et al., 2012). This way of inducing talker variability may not fully represent differences found in actual talkers (Pollack et al., 1954; Watkins, 1991). A more recent study showed that a change in talker can significantly diminish spectral contrast effects and thus challenges this notion that talker information plays no role in spectral contrast effects (Assgari & Stilp, 2015). These experiments

added either a +5 dB or a +20 dB spectral peak to sentences from one talker versus 200 different talkers. No significant difference was observed between conditions when filter gain was +20 dB. However, when a +5 dB peak was added, there was a significant decrease in the size of the contrast effect when sentences were produced by 200 different talkers. Therefore, there seems to be an increased sensitivity to the number of talkers when modest peaks are added.

The current study hopes to investigate the number of talkers needed to significantly reduce the contrast effect. Assgari & Stilp (2015) used 200 talkers and 200 sentences to introduce variation in speech. This large amount of variation may not be necessary to diminish spectral contrast effects and such a variation may not be of practical significance. Other experiments of talker variability look at word recognition. A classic study by Creelman (1957) showed that there was decreased performance in word recognition as the number of talkers increased from one, two, four, eight, and sixteen talkers. This study will likewise look at one talker, four talkers, eight talkers, and sixteen talkers speaking 160, 40, 20, and 10 sentences respectively. Using a modest, more natural, +5dB filter gain, it can be predicted that spectral contrast effects will be diminished with increasing talker variability.

Method

Participants

Nineteen Psychology students at the University of Louisville who are at least 18 years old, are native English speakers, and have no known hearing impairments participated in this study. The participants were compensated by course credit.

Stimuli

Sentence Selection. Sentences for the stimuli were drawn from a collection of recordings of high-quality speech called the Buckeye corpus. The corpus contains conversational speech

from 40 different speakers both male and female (Pitt et al., 2007). Apart from the one-talker condition, which was selected to be a male, all other conditions roughly balanced gender (half males and half females were used in the four-talker and sixteen-talker condition while the eight-talker condition had three males and five females). Pitch was not controlled in the experiment.

After talker and condition selection, a sample of sentences were drawn from each talker in excess of the number needed for each condition using Praat (Boersma & Weenink, 2016). Each sentence's long term energy spectrum was analyzed to see the energy naturally present in the low F_1 and high F_1 range. A Matlab script that integrated over these frequency regions produced a quantitative measure of this energy. Final sentences were chosen based on the least difference between the energy at the low F_1 range and the high F_1 range (see Table 1 for sentence duration and pitch). In other words, preference was given to sentences with a relatively flat long term average spectrum (final sentences ranged from -3.62 dB to +8.47 dB difference). Looking at the sentences in this way ensured that sentences did not have large energy differences prior to filtering, which would bias the results.

Stimulus Processing. The experiment consisted of four conditions: one talker with 160 sentences, four talkers with 40 sentences each, eight talkers with 20 sentences each, and sixteen talkers with 10 sentences each. Half of the sentences from each talker were filtered to emphasize the low F_1 frequency range (100 – 400 Hz) while the other half of the sentences were filtered to emphasize the high F_1 range (550 – 850 Hz). This was done by adding a +5 dB peak to these regions using the fir2 function in Matlab with 1200 coefficients (see Assgari & Stilp, 2015). Due to a small amount of ambient noise, sentences from one of the talkers in the four-talker condition were notch filtered at a 60 Hz center frequency (from 45 – 75 Hz).

Vowels. Target vowels were the same ones used in Assgari & Stilp (2015) and ranged on a ten-step continuum from /ɪ/ to /ɛ/ (refer to Assgari & Stilp, 2015 and Winn & Litovsky, 2015 for details on vowel production). The continuum allows for some vowels to be expressly either /ɪ/ or /ɛ/, while others are in between these endpoints and thus ambiguous. Target vowels were appended to sentences following a 50-millisecond gap.

Procedure. The experiment was run in a sound attenuated booth using a Matlab script that guided participants through each trial of all four conditions. The stimuli were presented binaurally through circumaural headphones (see Assgari & Stilp, 2015 for details on all equipment used). Each condition was presented to all participants in a random order and all trials within a condition were also randomized. Participants judged the vowel to be either "'eh' as in bet" or "'ih' as in bit".

Results

The percent /ɛ/ responses for each of the two types of sentences (those with either low F_1 or high F_1 emphasized) were used as the dependent measure (Figure 1). If a spectral contrast effect is observed, we would predict that for low F_1 sentences, there would be higher percent /ɛ/ responses. The difference between percent /ɛ/ responses for the low versus high F_1 condition measures the magnitude of the spectral contrast effect.

Responses to the end points of the /ɪ/ to /ɛ/ continuum were examined to ensure participants could identify endpoints as intended categories. Only data from participants that were able to differentiate these non-ambiguous sounds at least 80% of the time in every condition were considered. Nine out of the nineteen participants did not meet this criterion, leaving a sample size of 10.

An ANOVA test was not used because of the increased likelihood of type II error rate when multiple group means are highly similar. Instead, one-sample t-tests were used to see if each condition differed from zero. The one-talker (M = 0.040, SD = 0.067, t_9 = 1.89, p = 0.046), four-talker (M = 0.064, SD = 0.064, t_9 = 3.13, p = 0.006), and eight-talker conditions (M = 0.059, SD = 0.039, t_9 = 4.80, $p < 0.001$) all showed statistically significant spectral contrast effects. The sixteen-talker condition (M = 0.024, SD = 0.053, t_9 = 1.40, p = 0.097), however, did not show a contrast effect that was significant. This suggests that contrast effects are diminished somewhere between eight and sixteen talkers.

A paired sample t-test was performed between the one-talker and four-talker (t_9 = -0.72, p = 0.487), four-talker and eight-talker (t_9 = 0.27, p = 0.791), and eight-talker and sixteen-talker conditions (t_9 = 2.19, p = 0.057). The general trend (Figure 2) from the four-talker condition to the sixteen talker condition leans towards a reduction in contrast effect but no definite conclusions can be drawn since the difference was not significant. However, the eight-talker versus sixteen-talker comparison approached significance. These comparisons are consistent with a diminished contrast effect between eight and sixteen talkers when testing against zero effect as above.

Discussion

Spectral contrast effects have been shown in a variety of contexts. Although many studies have described this effect as a general acoustical phenomenon, reports of a diminished contrast effect by Assgari & Stilp (2015) call into question the influence of higher-level factors like talker variability. The purpose of the present study was to expand these findings and investigate what number of talkers is necessary to diminish this effect.

The results of the experiment show that there was a significant contrast effect for all conditions except for the sixteen-talker condition, which failed to produce a contrast effect. This suggests that spectral contrast is influenced by talker variability, such that as the number of talkers increases there is a decrease in contrast effect. However, the difference between any two conditions was not significant. Despite this, a sizable drop-off is noted when going from eight talkers to sixteen talkers. Considering the eight-talker condition produced a contrast effect and the sixteen-talker condition did not, it is reasonable to predict that these two points can be differentiated with more statistical power (especially since the current study uses ten participants and there is a lot of within-condition variability). Thus, there seems to be a diminishing spectral contrast effect somewhere between eight and sixteen talkers.

The reduction in spectral contrast effects can be explained by the fact that a change in talker requires a readjustment by the listener to stable spectral properties of the new talker (see Assgari & Stilp, 2015). In terms of sensory adaptation, frequent change in talker does not give our sensory system enough time to adapt to a specific talker's stable characteristics. Without this adaptation, the sensory system cannot regard the spectral properties of the new talker as reliable and thus a smaller contrast effect is observed, if one at all. In other words, change in talker obscures what can be considered reliable stimuli by our sensory system.

In the case of the one-talker condition, the size of the contrast effect appears to be low in comparison to the four- and eight-talker condition. This result was heavily influenced by two participants that showed a reverse contrast effect. This can be attributed to differences in how those individuals categorize vowels. A potential reason for this difference can be the language background of those participants. For example, in Sjerps & Smiljanić's (2013) multilingual study, the authors found that "overlap across various vowel categories in the F_1-F_2 space may be

reduced in Spanish" (p. 209). It is important to note that the present study required native English speakers but participants were not screened.

The current study is different from previous studies of talker variability and spectral contrast in that the study used (1) up to sixteen talkers, (2) modest +5 dB filter gain, and (3) conversational speech from the unique talkers. These three factors can help reconcile the seemingly contradictory findings of Laing et al. (2012) and the present study. Laing et al. (2012) used one talker and adjusted the formant frequencies to make it appear as if four talkers were used. This approach does not fully encompass true differences in talkers (Pollack, 1954; Watkins, 1991). The present study found the effects of talker variability approaching significance between eight and sixteen talkers and virtually no difference between effects from four and eight talkers. This latter finding is consistent with reports from four "talkers" used by Laing et al. (2012). No effect of talker variability on spectral contrast effect was found using +20 dB peaks (Assgari & Stilp, 2015). Since Laing et al. (2012) used large filter gain, we would not expect to see an influence of talker.

Findings from the present study also show that spectral contrast effects are more robust against talker variability when compared to studies that measure performance on word recognition tasks (Creelman, 1957). These types of studies on talker variability show a decreased performance with as little as two talkers, while this study shows contrast effects are present up to eight talkers (Creelman, 1957).

Using conversational speech is not a convention in studies of spectral contrast effects. A lot of variability can be conveyed in this form of natural speech when compared to speech designed for lab use. Conversational speech has a high degree of variability in terms of differences among individuals in coarticulation of sounds, individual differences arising from

differences in anatomy of the vocal tract, social and cultural differences in speech production, and differences in emotional state (Mitterer, 2006). This component of the present study thus extends the spectral contrast effect towards a more natural setting offering a higher degree of external validity.

Although conversational speech increases external validity, it does not point to any specific characteristic of talkers that may be contributing to the observed reduction in spectral contrast effects. Future studies should manipulate different spectral properties of speech that vary across talkers to differentiate the influences of these individual properties. This should be done in combination with smaller filter gains in order to increase sensitivity to these subtle influences on spectral contrast effects.

In conclusion, spectral contrast effects can be influenced by talker variability (Assgari & Stilp, 2015). This study shows that the amount of talker variability need not be as large as 200 talkers, but rather a diminished effect can be observed with as few as sixteen talkers.

References

Assgari, A. A. & Stilp, C. E. (2015). Talker information influences spectral contrast effects in speech categorization. *J. Acoust. Soc. Am. 135*(6), 3023-3032.

Barlow, H. B. (1961). "Possible principles underlying the transformations of sensory 597 messages," in Sensory Communication, edited by W. A. Rosenblith (MIT Press, Cambridge, MA and John Wiley, NY), pp. 53-85.

Boersma, Paul & Weenink, David (2016). Praat: doing phonetics by computer [Computer program]. Version 6.0.1, retrieved 04 January 2016 from http://www.praat.org/

Creelman, C. D. (1957). Case of the unknown talker, *J. Acoust. Soc. Am. 29*, 655.

Holt, L. L. (2005). Temporally nonadjacent nonlinguistic sounds effect speech categorization. *Psych. Sci., 16*(4), 305-312. 548

Holt, L. L. (2006b). The mean matters: Effects of statistically defined nonspeech spectral distributions on speech categorization. *J. Acoust. Soc. Am. 120*(5), 2801-2817.

Joos, M. (1948). Acoustic phonetics, *Language*, 24 , 5–136.

Ladefoged, P. & Broadbent, D. E. (I957). Information conveyed by vowels. *Journal of the Acoustical Society of America, 29*(1), 98-104.

Laing, E. J., Liu, R., Lotto, A. J., & Holt, L. L. (2012). Tuned with a tune: Talker normalization via general auditory processes. Front. Psychol., 3, 203. doi:10.3389/fpsyg.2012.00203

Mitterer, H. (2006). Is vowel normalization independent of lexical processing? *Phonetica, 63,* 209-229.

Pitt, M.A., Dilley, L., Johnson, K., Kiesling, S., Raymond, W., Hume, E. and Fosler-Lussier, E. (2007) Buckeye Corpus of Conversational Speech (2nd release)

[www.buckeyecorpus.osu.edu] Columbus, OH: Department of Psychology, Ohio State University (Distributor).

Pollack, I., Pickett, J. M., & Sumby, W. H. (1954). On the identification of speakers by voice, *J. Acoust. Soc. Am. 26*, 403-406.

Sjerps, M. J., & Smiljanić, R. (2013). Compensation for vocal tract characteristics across native and non-native languages, *Journal of Phonetics, 41*, 145-155.

Stilp, C. E., Alexander, J. M., Kiefte, M., & Kluender, K. R. (2010). Auditory color constancy: calibration to reliable spectral properties across nonspeech context and targets. *Atten. Percept. Psychophys., 72*(2), 470-480.

Stilp, C. E., Anderson, P. W., & Winn, M. B. (2015). Predicting contrast effects following reliable spectral properties in speech perception. *J. Acoust. Soc. Am. 137*(6), 3466-3476.

Watkins, A. J. (1991). Central, auditory mechanisms of perceptual compensation for spectral-envelope distortion, *J. Acoust. Soc. Am. 90*(6), 2942–2955.

Winn, M. B., & Litovsky, R. Y. (2015). Using speech sounds to test functional spectral resolution in listeners with cochlear implants, *J. Acoust. Soc. Am. 137*(3), 1430–1442.

Table 1

Average Sentence Duration and Pitch

Condition	Group Duration Avg (s)	Group Pitch Avg (Hz)
One-talker	1.843	103.495
Four-talker	1.943	141.660
Eight-talker	1.842	144.260
Sixteen-talker	1.723	136.495

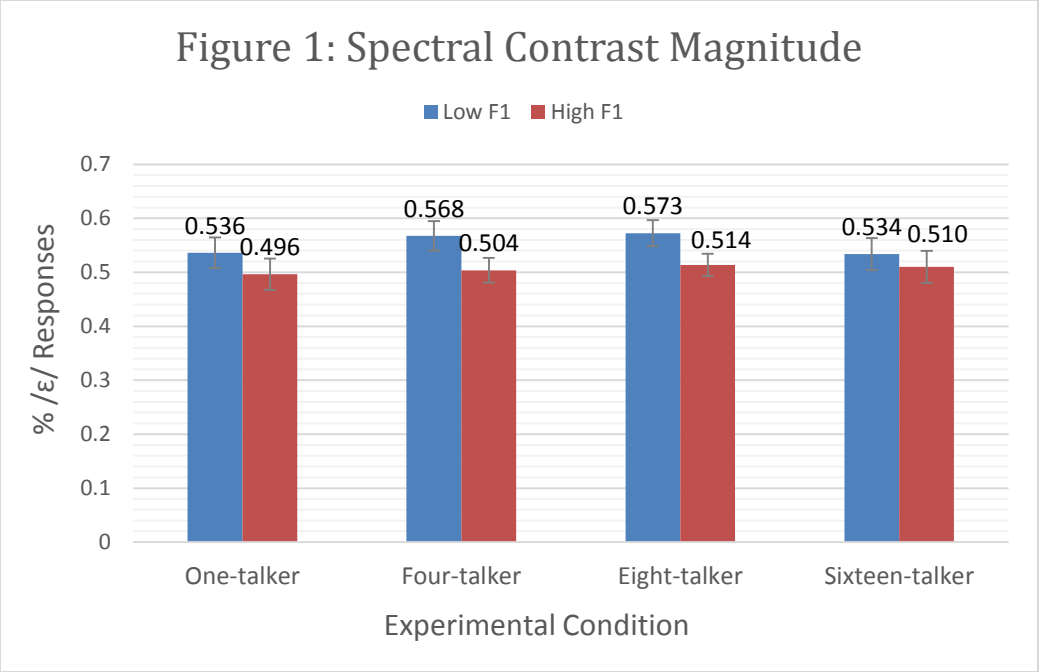

Figure 1. This shows the percent /ɛ/ responses for low vs high F_1 with standard error bars.

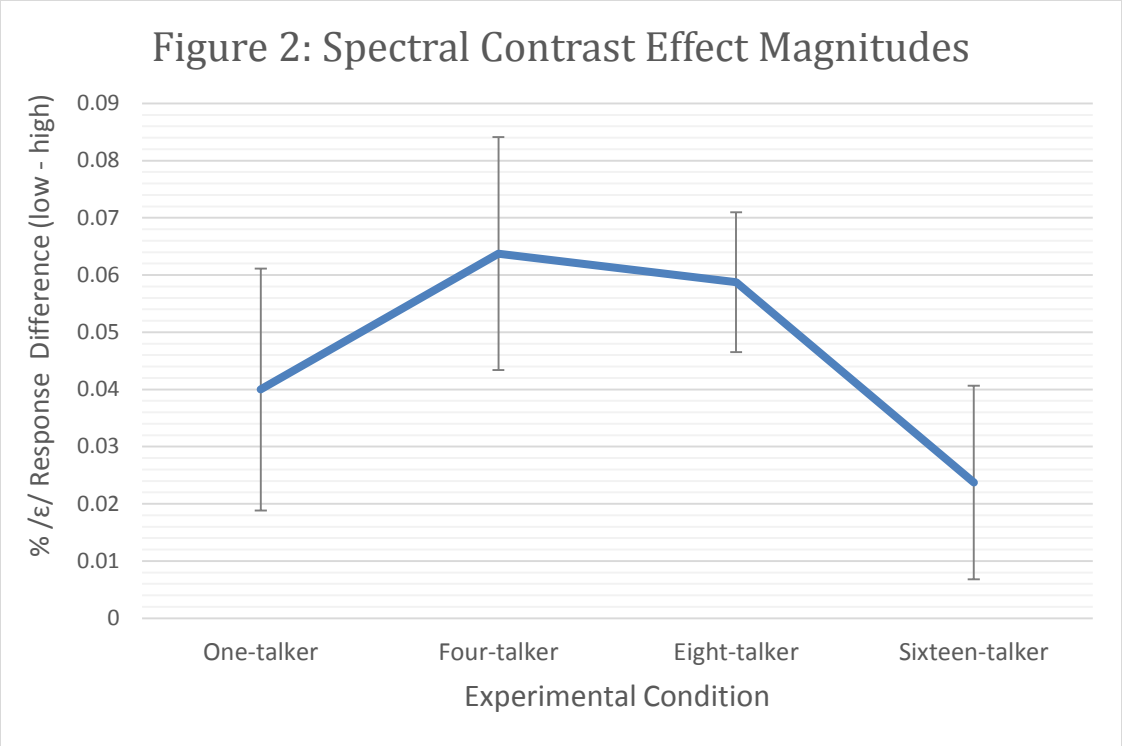

Figure 2. This shows the difference between the low and high F_1 condition with standard error bars.