The Acoustics of Shouting: A Case Study of English Vowels

Dine Mamadou

Rutgers University

The Acoustics of Shouting: A Case Study of English Vowels

Introduction

The current paper is an investigation of the acoustic properties of English vowels in shouted utterances; that is, we investigated whether or not vowel quality is affected in those utterances. Shouting is very demanding on the vocal tract and as such may cause the latter to undergo distortions of its shape. In this regards, the Source and Filter theory makes interesting predictions on the potential effects of yelling on vowel quality. Acording to this theory, speech sounds are distinguished on the basis of both the source and filter properties of the vocal tract (Maddieson 1984, Diehl 2008). That is, different configurations of the vocal tract and the activity of the glottis as the source will yield different vowel qualities.

So, if yelling has a distorting effect on the shape of the vocal tract, then the resulting sound will reflect the changes in the shape of the filter meaning the vowels under observation would display qualities that differ from vowel sounds that are normally uttered. In fact, Huber & al (1999) found that higher vocal intensities (a correlate of loudness) are typically produced with an increased jaw opening with a co-occurring decreased tongue height which in turn results in an increased F1. This predicts that vowels in shouting will be higher than those in normally uttered speech. Additionally, we can further predict that the transitions of the articulators in the course of a shouted utterance may take more time than they do in normal utterance; yielding longer vowels in the former condition than in the latter.

Ladefoged & Johnson (2011:92), point out that in English, "the first part of the diphthong is usually more prominent than the last. (. . .) the diphthongs often do not begin and end with any of the sounds that occur in simple vowels." Assume that the prominence of the first part of the diphthong is due to its position in the articulation chain, the higher the energy involved in an articulation process, the more peripheral the sound is likely to be. If this correlation is verified, the vowels involved in shouted diphthongs will have formant values that are closer to those of the monophthong versions of those vowels.

Building on the predictions of the discussion above, we hypothesize that:

H1: Vowel F1 will crease as a functon of intensity and

H2: Duration will be a good predictor of shouting.

Participants

Two native speakers of English (1 Male and 1 Female) from Iowa and Illinois, respectively were recorded. The female participant is a monolingual while the male participant speaks German (at an intermediate level) in addition to English. Both were informed of the comaprative purpose of the study but knew very little to nothing about the different points of comparison. The participants were recorded in a soundproof booth using **PMD661MKII Handheld Solid State Recorder**. The detachable microphone of the recorder was mounted to the participants' heads to ensure the distance between their mouth and the microphone is kept consistent at all time during the recording sessions. The recordings were done in mono and outputed in .wav format. In the shouted condition, the loudness threshold was set twice as high as the loudness threshold of the normal condition to prevent the higher frequences from being cut off.

Experimental design

The participants produced 16 randomized target words, comprised of 10 monophthongs and 6 diphthong in **hVd** context (an adaptation from Yoon & al., 2012). The 16 target words contain the monophthongs [i, ɪ, ɛ, ɑ, ɔ, u, ɚ, ʌ, æ, ʊ], and the diphthongs [eɪ, ɔɪ, oʊ, ju, aɪ, aʊ]. Shwa was not included because it we could not find a word with schwa in the hVd context.

The items were produced in two conditions: "normal" and "shouted". In the former condition, they were instructed to speak the way they would normally speak in a conversation with a person right in front of them, while in the latter, the instruction was to speak as though they were talking to a person who is about 100m away from them.

Each token was repeted twice in each condition, yielding a total of **128 target**

tokens. The target words were pronounced in isolation, as opposed to in a frame sentence, in order to minimize tiredome of participants, especially in the shouted condition.

Measurements and Material

F1, F2, Intensity and vowel duration measurements were harvested in Praat (version 6.0.32) using a script by Mietta Lennes (version 4.7.2003). F1, F2 and intensity measurements were taken at the midpoint of the middle 1/3 (at the steady state) of the vowels. Vowel duration was measured from the first zero crossing of the first harmonic of the vowel to the last zero crossing of the last harmonic before the stop closure (determined simultaneously on the basis of the waveform and the spectogram). Because each target word had a voiced coda, which have a robust second formant, the onset of the coda consonant closure was used as the right interval since it is more reliable than the offset of F2 in this case. Only the monophthongs are reported here.

The praat TextGrid was annotated in three tiers, the Formant tier, the Intensity tier and the duration tier. Each interval on each tier was encoded for the Condition (Normal and shouted) Vowel, Gender (Male and Female), Repetition (First = 1, Second = 2).

Data Analysis

The collected data was analysed in R (Version 3.4.3; R Core Team, 2017). Two sets of nested generalized linear models were fitted using the R function **glm()** with the Gaussian distribution family and "identity" as the link; under the assumption that the data is normally distributed. This assumption was later confirmed by eyeballing the normality and homoskedasticity of the plots of the variables. In order for our models to reflect our hypotheses, F1 was set as the criterion in one set of nested models and Duration in the other.

The predictors were both continuous (Intensity, (F1) and (duration))[1] and categorical (gender). Models with the same criterion were compared via the **anova()** function. In

[1]When F1 is the criterion, it's removed from the predictors; same for duration.

models with main effects and interaction, R^2 was obtained using the function **r.squaredGLMM()**. Experiment-wise alpha was set at **0.05**.

Results and Discussion

A descriptive summary of the data is presented in *Table1* below. F1 is in average *140Hz (214.89HZ)* higher in the shouted condition than it is in the normal condition, while intensity showed to be *5Hz* more in the former than in the latter. As far as duration is concerned, vowels are almost twice as long (*27ms vs 43ms*) when shouted than when normally spoken.

condition	meanf1	sd.f1	mean.int	sd.int	mean.dur	sd.dur
Normal	598.47	214.89	75.46	2.56	0.27	0.08
Shouted	740.90	198.32	80.06	4.53	0.43	0.15

Table1: This table shows the descriptive statistics of the data.

In the nested model with F1 as the criterion, a main effect of Gender was found ($p<0.05$) with an *R^2* of *0.18* (R^2 of centered measurements). Intensity was almost significant ($p=0.08$). More importantly, however, an interaction between gender and intensity was found ($p<0.05$) and the variance explained is *R^2=0.28*. That is, F1 increases as funtion of intensity based on gender.

In the nested model with duration as the criterion, gender was also found to have a main effect ($p<0.05$) with a variance explained of *0.18* (again in centered values). An interaction was also observed between intensity and gender ($p<0.05$) with an R^2 of *0.54*. The two nested models indicate that intensity is a good predictor of F1 and duration only if modulated by gender. More precisely, intensity is a good predictor of F1 in males while it only correlates with duration with females.

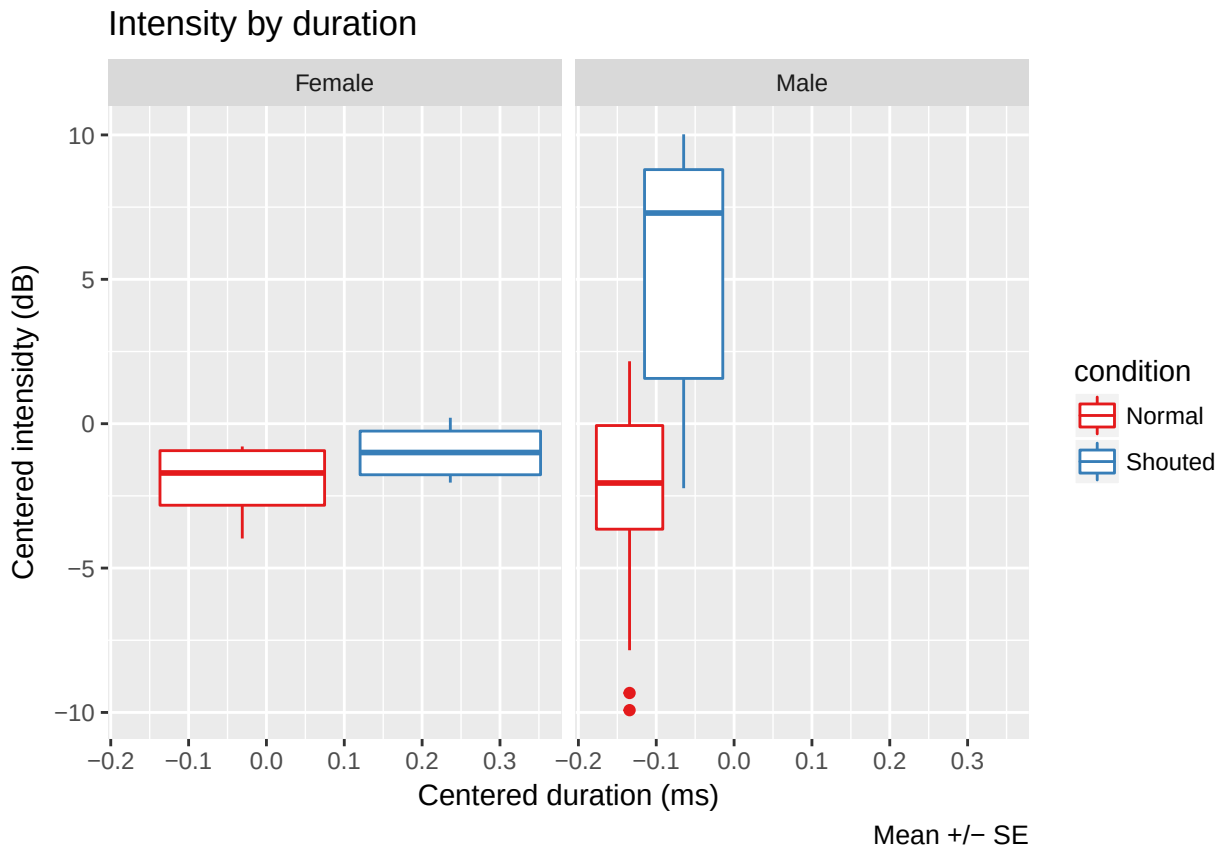

Figure1: This figure shows Intensity as a function of Duration by gender. In the shouted condition, Female vowel length is about 0.25(centered ms) while male's is only -0.12. Male compensate this by having a higher intensity value (7.5 in centered ms) while female's is slightly under o.

In both sets of models, the models with interaction offered the best fit for the data. While this did not exactly validate neither of our hypothesis, it did show that intensity alone is not a good predictor of vowel height (F1). This finding also partially aligns with Huber & al. (1999) in that only the Male speaker have higher vowels as a function of an increase in intensity, on one hand. On the other, the Female speaker does not use intensity as a cue for loudness, she uses vowel duration instead (*See Figure2 below*).

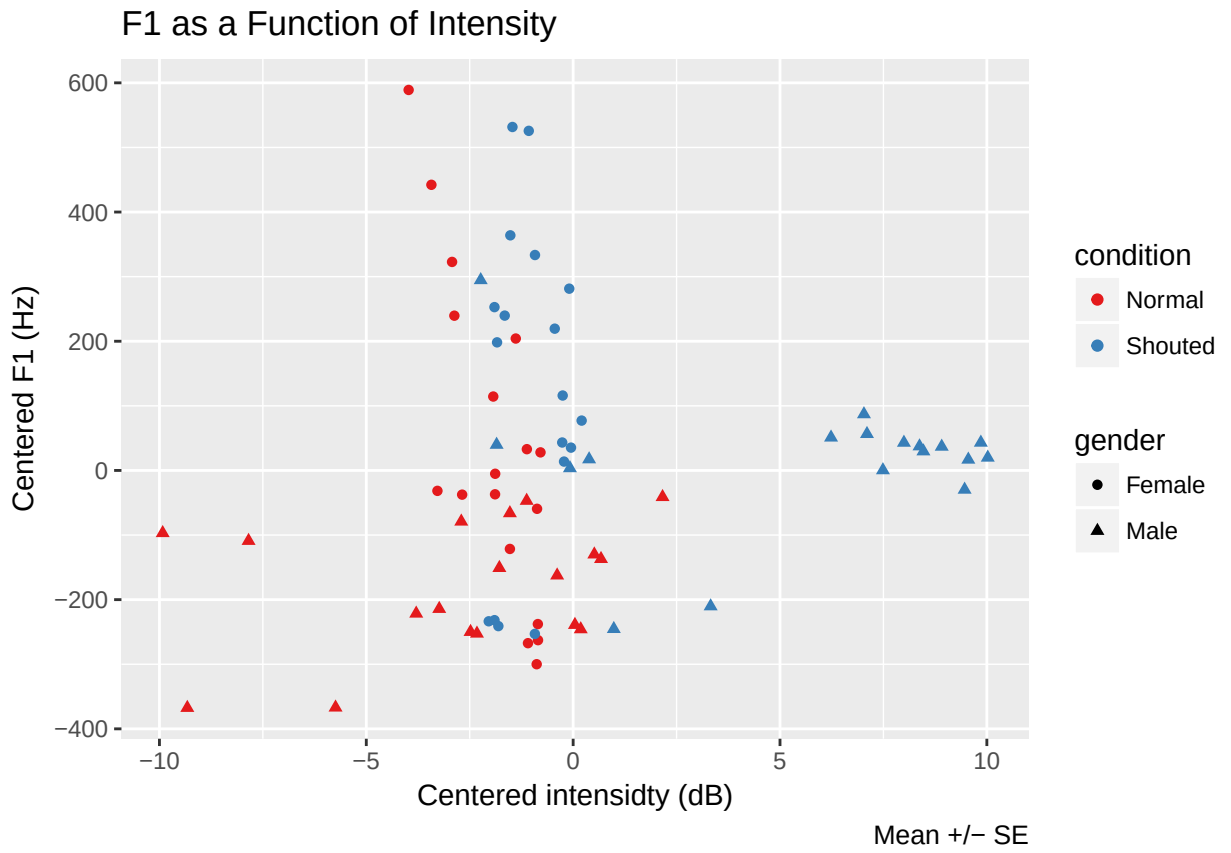

Figure 1

Conclusion

This paper investigated whether or not there's a difference in vowel quality between shouted and normally spoken speech in Female and Male. The findings supported the intution that there is indeed a difference. However, unlike a direct correlation between vowel heigh (F1) and intensity as suggested in the literature (namely Huber & al. (1999)), we found a more nuanced correlation. That is, Male speakers use intensity as a systematic cue for loudness and showed an increase in F1 as intensity increased. The Female speaker on the other hand systematically used vowel duration as a cue for loudness and only secondarily used intensity. While this systematic difference may reflect sociological and gender approaches to loudness, it highlights the importance of a gender sensitivity approach to these kinds of studies and constitutes a base on which future work can build.

References