A Proposed Framework for Rigorous Voice-Over Quality Evaluation

TJ Hatch[1]

Rensselaer Polytechnic Institute NY 12180, USA
hatcht@rpi.edu

Abstract. Rigorous measures of quality can sometimes be applied to many art forms including writing and drawing, but have not been developed for voice acting. By sampling listener opinions about various line reads and analyzing the acoustic profiles of those line reads, we arrive at an understanding of general acoustic patterns to follow in order to create a given effect.
This paper substantiates this concept on a small scale by following using line reads taken from Stephen Merchant's portrayal of Wheatley (Portal 2), and reads by an untrained actor that have not yet been procured. We find correlations between a read's perceived quality and its median frequency, cadence, pitch variance, and amplitude variance–and then illustrate how these properties may contribute to an understanding of an "ideal" read of the given line.
It is expected that quality correlations will not be strictly linear; they may more closely resemble a parabola, with a sweet spot and falloff on either side.

Keywords: Voice Acting · Vocal Expression · Voice Training.

1 Introduction

1.1 Motivation

Rigorous quality analysis of a line read has long been thought to be intractable on the basis that perceived quality of a line read is purely subjective. Where drawings can be measured based on how well the underlying forms match 3D space (see fig. 1) and works of fiction can be judged on grammar and internal consistency, it is hard to label a line read "correct" on any basis other than the presence or absence of microphone peaking, plosives, and audible breaths.

These measurements, however, don't represent an understanding on the part of the actor on how emotions manifest in the voice–they instead represent the presence of proper equipment and, in some cases, editing. Perhaps a better metaphor, then, would be diction. It is possible to imagine a system of practice where one reads a script and marks whether or not they pronounced each word correctly. Combining the number of correctly pronounced words with the multiplicative

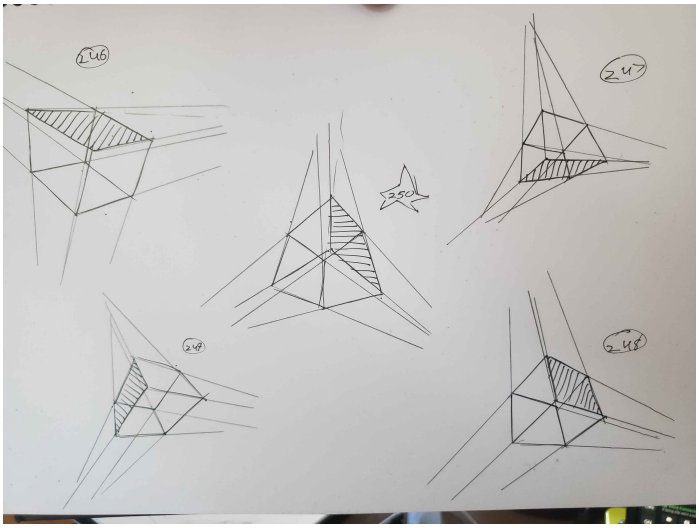

Fig. 1. A group of practice boxes. The lines on each box are projected backward to show whether they converge on the vanishing point–the better they converge, the more correct the box.

inverse of the time taken (in seconds) to read the script would yield a diction score which could conceivably be used for practice. This approach still does not aid in the full understanding of emotional expression, or of the physical methods used to truly make a voice line sound good. Diction is one part of a much larger whole in this respect, and if a system of feedback can be designed for the whole, that feedback can aid immensely in practice.

1.2 Research Problem

The goal of this paper is to create a convincing proof of concept for a larger study, which would define a set of target values for a set of key acoustic attributes to create friendly-sounding line reads. Those acoustic attributes are: median frequency, cadence, pitch variance, and amplitude variance. Further contributions will include a basic analysis of the difference between amateurs and professional actors in general, and a method for procuring amateur line reads based on professional line reads without exposing the amateur to the professional's work.

In addition, this paper will put forward a set of applications for a more complete body of work generated by larger studies based on this paper. These applications will range from tools to new ways of thinking about voice acting and communicating requirements.

2 Literature review

Previous studies cover listeners' recognition of emotion in some detail, but rarely look at their perception of "quality" in the clips they listen to. These studies often find that specific acoustic properties correlate strongly with recognition of specific emotions [1], [2]. These studies often contribute tables or lists of acoustic properties that positively or negatively correlate with recognition of a given emotion. They find little to no difference in recognition rates between genuine expressions of emotion and play-acted expressions of emotion in actors and non-actors alike.

These studies do, however, find noticeable acoustic differences in the three groups. They find that acted expressions of emotion tend to exaggerate the markers of the given emotion for added effect. Further, they show that trained actors' expressions of emotion differ further from the genuine article than those of their untrained counterparts. A valid interpretation of this result is that the acoustic properties of a given emotion (say anger for the sake of argument) can be taken as a baseline. Actors and non-actors alike understand how to produce these sounds with their own voices. Trained actors differ from non-actors in their understanding of these acoustic features and in their ability to draw them out in speech. In much the same way a cartoonist exaggerates core features of a face to create the most expressive picture possible (more realistic than reality), a voice actor may operate on the same principle.

[3] explore "modes" of speech in their paper. They take a physiological approach to the task of categorizing vocal expressions of emotion. They found combinations of various vocal "modes" for several emotions in their work. For anger, "falsetto with raised larynx, close jaw; retracted body tongue and elongated lips" was one such combination. Another was "hyperfunction, harshness and vocal fry." The observation that specific oral positions create specific emotional affects should not be at all surprising. Further, the notion that different oral positions should produce sounds with different acoustic properties is very intuitive. This acts as further evidence that the idea to use acoustic properties of speech as a measure of expression quality has merit.

3 Method

The method for this study can be divided cleanly into four procedure phases, each of which is relatively simple on its own.

3.1 Professional Line Collection

Five lines were selected from a .zip file containing all of Wheatley's voice lines, downloaded from sounds-resource.com. These lines were chosen for their generally friendly and unassuming tone, which is a core marker of Wheatley's character.

To find the median frequency of a line, Audacity's Pitch Detect plugin will be used on all clips at .5 second intervals. The median of the pitch measurements across the given clip is the median frequency of the full clip.
The pitch variance of a clip is the total difference between the highest measured frequency in the clip and the lowest measured frequency in the clip.
The cadence of a clip will be measured in words-per-second and averaged across the entire clip to create one number per clip.
Amplitude variance will be measured as follows: each clip will be amplified to have a peak amplitude of 0db. The lowest peak amplitude found during speech of a word after this amplification (i.e. -6db) is taken as the amplitude variance of the clip.

3.2 First review round

Five people who have no prior experience with Portal 2 were instructed to listen to these lines and fill out a form with the following questions about each:

1. What emotion was the speaker attempting to convey? (short answer)
2. How easy was it to recognize this emotion from their voice? (1-5 scale)
3. Did any parts of the line stand out to you compared with the rest of the line?
4. How genuine (i.e. realistic, non-acted) does the line sound? (1-5 scale)

They listened to the Wheatley lines and filled out the forms on their own devices.

3.3 Amateur Line Recording

An amateur (i.e. non-actor) was be brought into a quiet recording environment and given a transcription of Wheatley's lines. For each line, they were given all listener responses to the qualitative questions regarding the specific line as directions to guide their acting approach. They were allowed to record as many takes as they felt necessary to arrive at one they felt comfortable submitting. Lines were recorded on an Ars Tecnhica ATR2100x.

3.4 Second review round

Two more groups of five were be instructed to review lines following this recording session. One group reviewed only the amateur lines, and another reviewed both the amateur lines and the professional lines back to back. All reviews used the form described above. Participants were vetted prior to filling out the form to ensure that they had not played Portal 2.
The lines used are listed below:

1. "Hello! This is the part where I kill you!"
2. "Ah! There you are! Great, let me just get rid of this catwalk."

3. "'Lair' - heh, weird isn't it? First time I've said it out loud. Sounds a bit ridiculous, really. But I can assure you it is one. A proper lair. Deadly lair."
4. "I'll tell you, if I was up against impossible odds, this is the way I'd want to go out. Mashed with dignity. That'd be the way I'd choose."
5. "Designed this test myself. It's a little bit difficult."

4 Results

4.1 Acoustic Measurements

	Cadence (words per second)	Average Pitch (Hz)	Pitch Variance (Hz)	Amplitude Variance (dB)
1	3.581	216.6	385	-10.16
2	2.683	189.625	135	-11.098
3	3	169.455	119	-7.804
4	2.764	135.143	63	-13.739
5	2.668	157.4	78	-3.255

	Cadence (words per second)	Average Pitch (Hz)	Pitch Variance (Hz)	Amplitude Variance (dB)
1	4.054	279.25	85	-4.826
2	3.77	251.333	150	-8.706
3	3.339	152.412	346	-4.17
4	4.319	161.077	176	-6.215
5	3.344	232.2	170	-0.296

Fig. 2. The first table is the list of acoustic measurements for the amateur actor, and the second is the list of measurements for the professional.

Several measurable differences were found between the amateur and the professional voice actor. The amateur spoke at a lower general cadence, and had a lower pitch variance overall. The amateur's amplitude variance was higher than the professional, as well. The amateur spoke at a lower overall pitch, often by around 60 hz.

4.2 Survey

12 unique people participated in the survey. In total, 82 survey responses were collected from these 12 people. Five people responded to the "group one" question battery, three people responded to the "group two" question battery, and four people responded to the "group three" question battery. In total, two more people rated the professional's voice lines than the amateur's.

Survey participants rated the amateur voice actor much more consistently on emotional readability (question 2) than the professional, hovering around 3.5/5

for all lines. Every readability score for the amateur fell between the readability scores for the professional's 2nd- and 3rd-most readable lines. The finding that this score is much more consistent than the professional's readability score is supported by Fig. 3 ($p<.05$ via F-test).

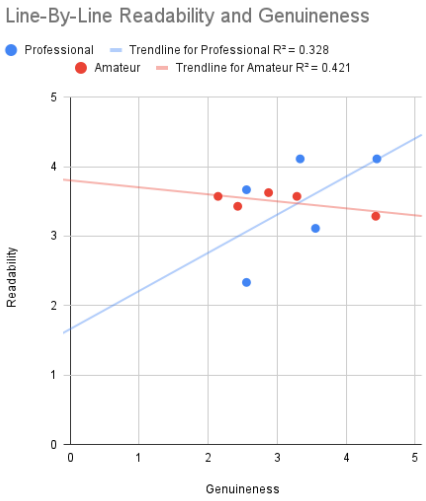

Fig. 3. Each point on this scatter chart represents the aggregate score of a voice line on questions 2 (y-axis) and 4 (x-axis)

No line averaged a score below 2 on any question in this survey. However, individual reviewers were willing to hand out low scores; reviewers gave a line 1 on a quantitative question 16 times in total, 10 times on question 2 and 6 times on question 4. Only one of the 82 survey responses had "one" for both questions. This response was with reference to the professional actor's reading of line 1. The respondent's answers to questions 1 and 3 are as follows: "an... ironic? joy?", and "obviously the part where they wanna kill me but with how fake the whole thing is it's hard to tell just what emotion they're actually putting behind it", respectively.

5 Discussion

The most significant finding from the surveys was that the amateur received very consistent genuineness scores compared to the professional actor. It is possible that this is because of the scripts themselves, and what those scripts may call for. For instance, if a script calls for the speaker to express sarcasm in their voice, it seems likely that a good actor would receive a lower genuineness score. It is also

important to note that the actor received higher peak genuineness scores. These points, taken together, suggest that the professional's more variant genuineness scores may be a good thing.

This result calls into question the strength of genuineness as a proxy measure for quality. It may be a good proxy in some circumstances, but it seems clear that not all situations call for genuineness.

This line of thinking also raises the question of whether it is useful to think of readability as a proxy for quality. Should a character's emotions always be readable? Do there exist scripts where a character's emotions ought to be difficult for a listener to discern? Do some scripts simply make it more difficult for a listener to discern the speaker's emotion? In cases where a character's feelings are complex or nuanced, it may simply be more difficult for a reader to parse everything out, regardless of how well the actor does.

The acoustic differences between the two actors could come from any number of places, and they could impact line quality in any number of ways. It is impossible, using only the data we have, to identify any true causes, but it is possible to create some hypotheses. For instance, it is of special note that the trained actor had higher pitch variance than the amateur. His pitch variance was at its lowest in his recording of audio clip 1, at only 86 hz. Audio clip 1 also has the lowest mean genuineness score. The relation goes further, with the second lowest pitch variance and genuineness scores (for the professional actor) belonging to his clip 2. The amateur's lowest pitch-variance clip is also his lowest-scoring genuineness score, but the relation for the amateur is much weaker in general than it is for the professional. This result may suggest a relation between pitch variance and genuineness; further information is definitely required, however.

6 Future Research

There are a great number of ways to expand this project. An easy way is to simply run the study with more data points. With more actors, amateurs, lines, and survey respondents, statistical relations will only become clearer, and conclusions will be easier to draw.

A future study could be further strengthened by improving the survey. This study only uses proxy measurements for quality, but by attempting to directly measure quality future studies would gain a very useful perspective on the strength of other subjective qualities of a line read (e.g. readability, genuineness). For instance, a genuineness score may sometimes match a quality score and it may sometimes directly disagree with it. By finding if or when this happens, we would learn a lot about how people perceive and think about voice acting, as

well as what they truly value in a line read. The same goes for a readability score.

Furthermore, there are a number of other potential proxy measurements for quality that could be just as, if not more, useful than the direct measurement. For instance, is there a difference between asking a user about their enjoyment of a line and the perceived quality of that line? Perhaps a line is good if it creates a clear image in the listener's head, either of a character or of a situation? The measurement of quality is a topic that requires extensive thought, care, and research to solve or understand, and different measures of quality may affect what results a study provides. It is useful to gain as comprehensive an understanding of this topic as possible.

It is also important to think about ways the acoustic measurements could be iterated upon. In this study, the acoustic measurements were global to an audio clip; they covered or gave information about the clip as a whole. Pitch was averaged across the clip, and pitch variance was based on the clip's highest and lowest moments. In order to arrive at a more comprehensive understanding of the art, it will be useful (and likely necessary) to evaluate the clip on a more granular basis.

One tool to address this is to generate and use pitch profiles for each line; by sampling a line at regular intervals for pitch and plotting those out over the course of the line, it would be possible to generate a profile (that would be easy to graphically represent) representing the curve. Then, by plotting data about where listeners perceive emphasis on a timeline as well, one could observe how that emphasis aligns with the pitch profile. It is even easier to do this with an amplitude profile, because amplitude profiles are automatically generated by most audio editing programs when audio is recorded or imported.

Another interesting experiment would be to normalize lines by different actors for fundamental frequency by pitching them up or down to match each other, and then running the same battery of measurements on them. There are, of course, questions about how best to perform this operation; pitching an audio waveform down by spreading its samples apart will naturally cause it to slow down, and other methods can sometimes create audio artifacts that are undesirable for testing. Work in this area would require a good understanding of different pitch shifting methods.

Future applications for more fleshed-out versions of this work are also readily apparent. For instance, a tool that takes in audio, generates a set of acoustic properties, and makes audience score from those predictions seems like a viable target to shoot for. This hypothetical tool could come with curated scripts, target pitch profiles, and data sharing capabilities for further improvements both among users and for the tool itself. It could also make use of LLM technologies to generate new scripts or target pitch curves on the fly at varying difficulty

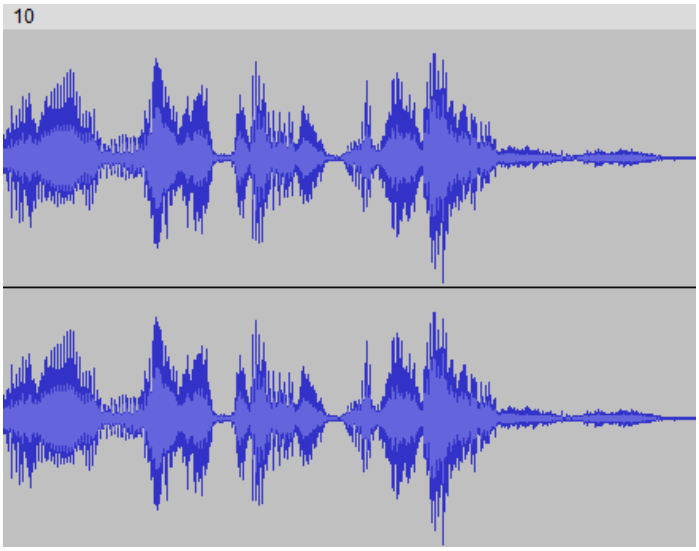

Fig. 4. A waveform of an audio clip, which is itself made up of samples taken at tiny fractions of a second. Any number of techniques may exist to smooth this into a continuous, readable amplitude profile–and similar techniques would be key to creating pitch profiles.

levels.

By bringing this type of measurement into the consciousness of actors, both amateur and professional, it may be possible to create new tools for direction and communication among actors. If target pitch profiles, for instance, were to find usage in the industry, we may see casting calls that include target pitch profiles in the audition directions. The discourse around how to characterize and arrive at voices may also take a more rigorous turn as findings solidify.

Another vector of study that could be taken in the future is the study of different archetypes, and of what qualities are desirable in each. If common tells could be found that indicate reliably whether a character is a villain, for instance, that knowledge would have a number of easy-to-imagine uses. One might avoid those tells to create a subversive character, for instance, or one might practice them in order to better understand how to sound evil. If, however, no consistent differences can be found between, say, heroes and villains, then it may suggest that the most important thing to focus on developing is one "core" voice that an actor can then develop on.

7 Evaluation

As a purely statistic-based study, this paper does not carry much weight. It contains one statistically significant finding which, while interesting, cannot carry the full task of using acoustic properties to find differences between good and bad voice acting. The sample sizes in most cases are too small to arrive at statistically significant data, and between the actors in particular it is unclear what differences stem from an experience differential and what differences are purely random individual differences that exist in the general public.

However, as a model of a larger study, the study absolutely succeeds. Difficulties with the study, as have been discussed above, include the small sample size, the proxy measurements for quality, and the acoustic measurements being entirely global to their audio clips. By identifying these issues, the study has revealed a number of considerations that a larger, more committal study can understand and design itself around.

In addition, the running of the study has revealed multiple avenues for extension beyond both its original concept/questions and its methods. The study did more than simply validate one approach to the problem it set forward: it raised the possibility of several approaches (that are not mutually exclusive!) to the same problem. Within the bounds of this study's scope, it did everything it needed to and more.

Disclosure of Interests. The author declares no competing interests.

Acknowledgments. Thank you to Tim, to my survey respondents, to my TAs, to my classmates, and to my professor, Neha Keshan.

References

1. Grass A. Drolet M. et al. Jürgens, R. Effect of acting experience on emotion expression and recognition in voice: Non-actors provide better stimuli than expected. *J Nonverbal Behav 39*, 2015.
2. Russo Frank A. Livingstone Steven R., Choi Deanna. The influence of vocal training and acting experience on measures of voice quality and emotional genuineness. *Frontiers in Psychology*, 2014.
3. Madureira S. Viola, I. Voice qualities and speech expressiveness pontifícia universidade católica de são paulo. 2007.