

# **Consensus and Disagreement in the Interpretation of Electronic Voice Phenomena: The Ferry Plantation House EVP Project**

## **Summary**

**Mark Leary**

**Background:** Investigators often disagree regarding the interpretation of electronic voice phenomena (EVP), raising questions about the likelihood that any particular investigator's interpretation is "correct." Not only does an incorrect interpretation of EVP misinform investigators regarding the nature of paranormal activity, but reporting incorrect interpretations is misleading to clients and others.

**Goals:** This study had three main goals: (1) document the degree to which investigators agree or disagree on their interpretations of a set of typical EVP, (2) present a means of identifying the most likely interpretation of a particular EVP, and (3) determine whether multiple EVP recorded at the same location on different occasions contain any similar voices.

**The EVP:** To obtain a set of EVP for analysis, investigators were contacted who had conducted investigations at the Ferry Plantation House, Virginia Beach, VA. Over 250 EVP, all of them recorded without a background noise source, were initially submitted, from which 94 were chosen. These came from 11 investigators on 7 different investigations. In general, EVP were selected that had obvious vocal tone to make identification of similar voices easier. In addition, two clips that did not contain EVP were included to bring the number to 96.

**The Raters:** Twenty-four individuals (10 men, 14 women) agreed to listen to and interpret the 96 audio clips. The raters ranged in age from 29 to 62 (average age = 46). All but two currently belonged to a paranormal investigation group, and all but one had been on at least 1 paranormal investigation, with an average of 20 investigations each.

The raters were sent a CD with the 96 audio clips, along with an EVP interpretation form, a form for noting voices that sounded similar, and a background information questionnaire. The interpretation form asked raters to interpret each EVP, indicate any emotion that they detected in the voice, and rate their confidence that their interpretation of the EVP was correct. The background questionnaire asked about raters' beliefs in the paranormal and their interest in various paranormal shows, and included a brief measure of basic personality dimensions (such as extraversion, emotional stability, and agreeableness).

**The Consensus Interpretation:** For each EVP, a consensus interpretation was determined by identifying the words that were listed by the largest number of raters. Presumably, the most frequently nominated words reflect the interpretation that is most

likely to be correct. Put differently, the consensus interpretation of an EVP is more likely to be correct than the interpretation made by any particular person.

**Agreement among Raters:** With the consensus interpretation in hand, the percentage of raters who agreed with the consensus interpretation could be calculated. This number, which can range from 0% (no one agreed with the most common interpretation) to 100% (all raters agreed with the consensus interpretation), can be viewed as an index of both the clarity of an EVP and the degree to which independent raters agree in their interpretations of it. In this sense, it provides a more precise quantitative alternative to the traditional classification system in which EVP are classified as Class A, B, or C.

The EVP that had the highest agreement (“What’s going on?”) was listed by 83% of the raters. However, the overall agreement of the set of EVP was much lower. Across all EVP, average agreement with the consensus interpretation was 21%. In other words, only about 1 out of 5 raters listed an interpretation that agreed with the most common (and, presumably most probable) consensus interpretation. When analyzed at the level of particular words rather than the entire EVP, average agreement was 35% (that is, raters agreed with the most common interpretation of each word on about 1 out of every 3 words on average).

Some EVP not only had 0% agreement, but the various interpretations sometimes differed wildly. For example, one EVP that had no agreement across raters was interpreted as saying, among other things: Deep inside there’s a pickup; Keep those hidden Mr. Gel; He comes out here; Go outside and just lean on it; Get it tight, got to stretch it; Don’t try to persuade them; Get us out Mr. Kant; and I need the guns out if this is what you’ll do. Such interpretations do not even contain similar phonemes.

**Emotional Content.** Raters indicated whether they detected any emotion in the voice. The majority of the EVP (63.5%) had no discernible emotional tone. However, raters indicated that some EVP expressed sadness (9.7%), anger or irritability (8.2%), urgency (7.7%), and happiness (6.3%).

Setting aside the fact that most raters thought that most of the EVP had no emotional tone, when an emotion was detected, on average only 12.7% of the raters agreed that a particular emotion, such as anger or sadness, was present. Thus, raters showed even less agreement in detecting emotion than in interpreting the content of the EVP.

Raters differed notably in the degree to which they detected emotions in the voices. Interestingly, their tendency to hear emotions in the EVP was related to their own personalities. For example, raters who scored higher on the measure of extraversion reported “happiness” in the voices more frequently, raters who scored higher on agreeableness reported hearing both more “happiness” and more “anger/irritability,” those who scored higher on a measure of emotional stability heard more “happiness” expressed, and those who were higher in openness to new experiences detected more emotions overall. These data suggest that raters’ interpretations of emotional tone reflect their own personalities as much as the actual features of the EVP.

**Rater Confidence.** For each EVP, raters indicated how confident they were that their interpretation was correct on a 4-point scale (where 1 = not at all, 2 = a little, 3 = moderately, and 4 = very confident). Across all EVP, raters' confidence averaged between "a little" and "moderately" confident (mean confidence = 2.5 on the 4-point scale). However, confidence varied greatly across EVP, and raters indicated being very confident in many of their ratings.

Could raters tell which of their interpretations were correct? To examine this question, raters' confidence judgments were correlated with the number of their interpretations that agreed with the group's consensus (again, assuming that the consensus is more likely to be correct than any particular individual's interpretation). The correlation was .34, indicating a relatively weak relationship between raters' confidence in their interpretation and the degree to which their interpretation agreed with the consensus interpretation. Clearly, investigators cannot rely on their own subjective judgment of being correct.

**Differences among Raters:** Assuming that people whose interpretations agree more often with the group consensus are more likely to be correct than people whose interpretations differ from everybody else's, we can calculate an index of personal agreement that may tell us how good a particular rater is at hearing the most likely interpretation. Across the 24 raters, the raters agreed with the group consensus from 17% to 35% of the time, with an average of 22%. That is, the "best" rater agreed with the group consensus interpretations on 35% of the EVP, and the "worst" rater agreed on 17% of the EVP.

When analyzed at the level of the word rather than the entire EVP, percentage of agreement with the group consensus varied from 31% to 51%, with an average of 38% of the words. So, if we play the average EVP to a group of 100 people, only 22% will have the same entire interpretation, but 38% will agree on any particular word.

We looked to see if the characteristics of the raters mattered. The degree to which raters agreed with the group consensus interpretation was not related to the number of EVP that they had personally recorded, their years of involvement in paranormal investigations, to the number or content of paranormal television shows they watched, to basic personality dimensions, to their age, or to beliefs in the paranormal. The only variable that was significantly related to agreement was gender, with women's interpretations (mean = 24%) agreeing with the consensual interpretation significantly more than men's interpretations (mean = 20%).

Most raters gave interpretations that were meaningful phrases, but some gave phonetic interpretations even if they did not make semantic sense. For example, on one EVP for which there was no consensus, some raters gave meaningful interpretations (such as "Hey we sung in the chorus" or "That is so great, Cory"), whereas other raters wrote down what they heard even though it didn't make sense (such as "Hack me some green course" and "Hey peace and grin Coreys"). Investigators should consider whether imposing

meaning on an EVP may lead listeners to “hear” words that help the phrase make sense but that might be incorrect.

The raters also differed in their willingness to leave blanks. The rating form explicitly told raters to put an asterisk (\*) if they heard a word that they couldn’t interpret. Some raters used asterisks regularly, but others did not use them at all. Given that we can assume that no rater was perfectly confident of every word, those who filled in words that they didn’t understand probably made more misleading interpretations than those who admitted that they didn’t understand certain words.

**Similar Voices:** As noted, one goal of the study was to look for similar voices recorded by different investigators on different occasions. Raters listed EVP that sounded alike and wrote down what they had in common (for example, “EVP #s 4, 20, 47, and 96 sounded like a grumpy old man”). Across all EVP in the set, there were over 4,500 possible matches of voices, but only 9 pairs of EVP were identified as similar by more than 4 of the 24 raters. Clearly, the raters did not hear much similarity in the voices across the various EVP.

---

### **Conclusions and Recommendations:**

1. Investigators should have less confidence in their interpretations of EVP than they typically do. On average, the consensual interpretation was agreed on by only 22% of other people. And, of course, all interpretations other than the most common, consensual one were agreed on by even fewer people. In fact, most of the raters’ interpretations were not given by any other listener!
2. Furthermore, raters were not particularly good at judging the correctness of their interpretations. Thus, having the sense that “I’m sure this is what it says” does not indicate that other people will agree with one’s interpretation (or that it is actually correct).
3. In light of the fact that any particular investigator’s interpretation is not likely to be shared by other people and we know that people’s interpretations are biased by what they expect to hear, investigators should never interpret an EVP for other people without playing it for them several times and soliciting their independent interpretations (and maybe even not then).
4. If the interpretation of a specific EVP is particularly important, investigators should use a miniature version of the procedure used here. Have 10 people independently listen to the EVP and determine the consensus interpretation. Then report an interpretation of the EVP to others only if a majority agrees in their interpretation. Alternatively, investigators could report the % of people who agree with the most common interpretation. In some cases, it may be helpful to report more than one potential interpretation, along with the % of people who agreed with each one. Providing listeners with such data is a more honest and

responsible way to share EVP than to assert a particular interpretation that might, in fact, be quite idiosyncratic.

5. Investigators should be willing to refrain from offering interpretations of ambiguous EVP. Providing a questionable interpretation as if it is certain is misleading and sometimes dishonest. Just because an EVP can not be interpreted does not mean it is not a useful piece of evidence, so investigators should not feel compelled to interpret EVP that are unclear.
6. Because many investigators emulate the practices seen on popular paranormal shows, those shows should be careful not to offer confident interpretations of ambiguous EVP. When a consensus cannot be reached among the show's staff, the show should admit that the content of the EVP is unclear or present alternative interpretations of it. Alternatively, a show could post EVP from each investigation on a web site while the episode is in production, asking viewers to indicate what they think the EVP say. Then, the most common interpretations of each EVP could be presented when the show is aired. This practice would not only model responsible interpretations but would also avoid having some viewers dismiss the entire paranormal field because the show's interpretations do not reflect what viewers themselves hear.
7. Paranormal investigation groups should have formal guidelines for the interpretation of EVP that minimize the likelihood that their members will offer interpretations of EVP—whether to other members, clients, or outsiders—that are expressed with greater confidence than the objective evidence warrants.

The viability of EVP as useful evidence of paranormal activity depends on the degree to which investigators can trust their own and others' interpretations of them. Paranormal investigators should exercise greater care in offering interpretations of EVP, and procedures should be used to assure that clients, other investigators, and the public are not inadvertently misled regarding interpretations of the EVP that investigators record.

6/1/2010