University of Colorado Denver Analysis

The National Center for Media Forensics at the University of Colorado Denver conducted a forensic speaker comparison based on the industry-standard Likelihood Ratio (LR) analysis. In their test, they used the software package BATVOX from AGNITIO/NUANCE. The analysis compared voice and phoneme characteristics of two sets of samples: Sample A included five conversation segments published by the Joint Investigation Team in September 2016. Sample B included two conversations conducted by the Insider in the course of interviews with Gen. Tkachev in October and November 2017.

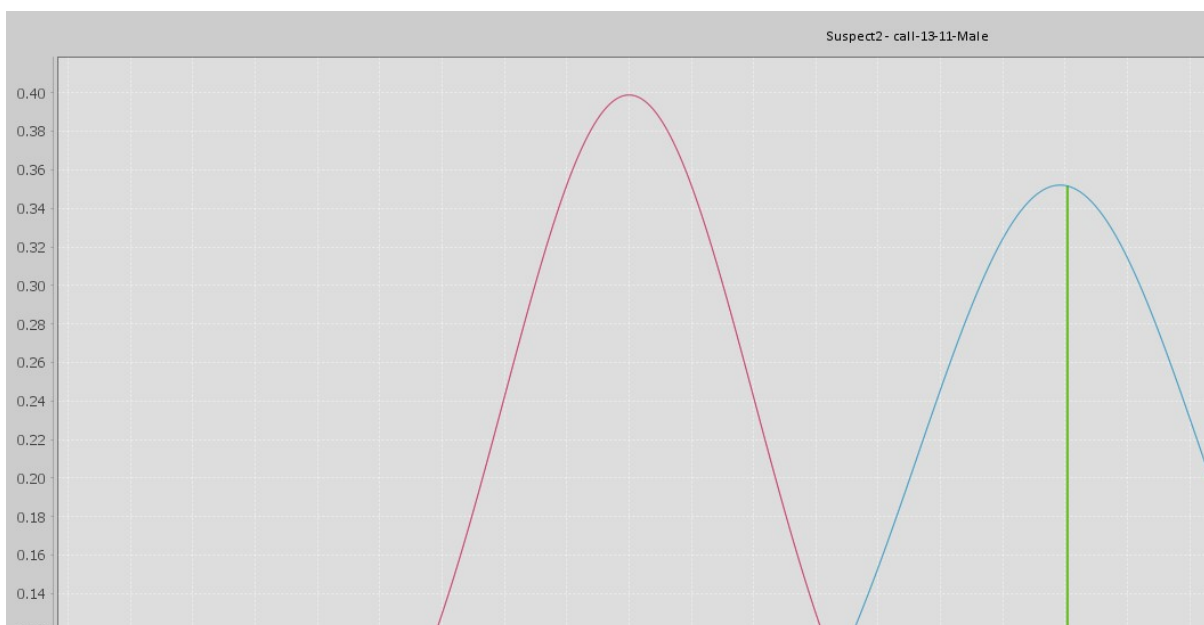The speaker comparison generated a Likelihood Ratio (LR) of 428.

Figure presenting the results of the automatic statistical analysis using BATVOX, the LR-value is stated top right. Provided by the National Center for Media Forensics at the University of Colorado – Denver. The red distribution corresponds to the comparison between different speakers, the blue distribution corresponds to the suspect/known speaker, the green bar corresponds to the evidence/unknown speaker and is inside the suspect's distribution.

The conclusion of the University of Colorado Denver speaker-comparison test was summarized as follows:

> "It is 428 more likely that the speaker Delfin is the same as the speaker in 'call_13-11-46_OUT_89122565174.aac' [*note: filename of the first call between The Insider and Nikolai Tkachev*], than the speaker Delfin is any other speaker. The analysis supports the hypothesis that the speaker Delfin is the same as the speaker in 'call_13-11-46_OUT_89122565174.aac'."

The value of 428 is a reference to the likelihood ratio of the two voices matching, providing what the European Network of Forensic Science Institutes would call "moderately strong support" for the hypothesis that Delfin and Tkachev are the same person behind the calls.

In real-life terms it means that it is 428 times more likely that voice from Sample A matches the voice from Sample B, than that it matches any randomly selected person from the reference universe. In this case, the reference population is defined as "Russian-male Speakers", and is based on a representative group of 42 different speakers (against a minimum recommended sample of 30 different speakers per group).

Prior to conducting the comparison, the UC-Denver system was calibrated using samples from 100 different Russian speakers recorded in normal conversational style under different quality and noise conditions, all of them with a Signal to Noise Ratio (SNR) greater than 15 dB, and between 30 seconds and 2 minutes net speech. During such calibration, two measures were obtained: first, an inter-variability LR (i.e., the likelihood ratio that any two *random* voices among this reference group of 100 belong to the same speaker), and second, the intra-variability LR (i.e. the computed likelihood ratio that any two voices *in fact belonging to the same speaker,* are the same. The inter-variability LR score was 78, while the intra-variability LR score was 181.

It is against the backdrop of this latter LR score, says Dr. Catalin Grigoras, director at National Center for Media Forensics, that the score of 428 should be assessed.

For reference, the voice segments from Sample B (Gen. Tkachev) were also tested against all male speakers in all other intercepted telephone conversations published by the Ukrainian Security Service (SBU) in the period 2014-2016. In these reference tests Tkachev's voice did not score a LR higher than 1 (i.e., no statistical relevance) with any sample except that of Delfin. As a reference test of what a definitive negative result would be in UC-Denver's analysis with BATVOX, a test was run comparing Tkachev's voice and the voice of The Insider's interviewer, producing a LR of 0.2043.

Dr. Grigoras stated that the LR from the BATVOX test must be evaluated in correlation with other evidence collected in the case, which may further increase its relevance. Based on a LR of 428, Dr. Grigoras exemplified this with the concept that if there were 428 potential randomly picked up suspects, a similar match would occur only once. If, however, the suspect population is restricted further, say to only Russian males aged over 50, holding a high military rank, and having the name and patronymic "Nikolai Fedorovich", the effective overall LR assessed by the investigators should rise sharply.

Another classic example to help readers understanding the relevance of a LR is to consider a crime in a village with 1000 inhabitants, and a LR pf 428 that one evidence found at the crime scene belongs to one of them and is a male. If we exclude the women, children, and all the males younger than 50 then the suspect population can be less than 200 and the hypothesis that the evidence criminal is one of them has a higher support.

Dr. Grigoras added that it is always highly recommended to correlate one evidence (in this case the automatic speaker recognition results) with some other forensic analysis (e.g. phonetics, fingerprint, DNA, etc.) or evidence (e.g. voice line ups, witnesses, GPS, etc.).

[Note: The original document provided by UC-Denver was modified by Bellingcat to make changes to the formatting and text style. The file has also been saved as a .pdf file, while originally being sent in the .rtf format. The content has not been changed.]