Automatic Voice Comparison Performance in Forensic Casework

Timo Becker[1]*, Gaëlle Jardine*[2]*, Yosef Solewicz*[3] *and Stefan Gfrörer*[1]
[1]*Federal Criminal Police Office, Germany*
{timo.becker|stefan.gfroerer}@bka.bund.de
[2]gaelle.jardine@gmail.com
[3]*Israel National Police*
solewicz@police.gov.il

Introduction

One of the variables usually given to express the performance of a forensic voice comparison system is its error rates. A court that is presented with the results of ex perts' voice comparisons needs to be sure that what is stated as a system's theoretical error rate also applies to the one particular case being presented. This is not always the case: theoretical error rates are usually based on evaluations with speech corpora (the most common one being NIST (Przybocki et al. 2007)) that generally are of very much higher quality than forensic speech samples and therefore are likely to give better results.

Experts who therefore prefer to conduct their own evaluations in order to calculate error rates that they can reasonably claim to be appropriate to their given case are often confronted with the impossibility of collecting a big enough evaluation speech corpus, especially if their case contains channel mismatch.

Some experts will go ahead with the automatic voice comparison anyway and publish the "theoretical" error rate. They will obviously have to alert the court about the fact that in their specific case, at least if it contains typically forensic-quality speech samples, the actual error rate is unknown, but almost certain to be higher than the one being advertised. Other experts will prefer to avoid this uncertainty and choose not to use the system at all. In this case, however high-performing the system may be in theory, i.e. however low its theoretical error rate may be, in practice the system, not being used at all, has zero performance.

We would like to address this problem of real rather than theoretical performance and suggest a new way not only of defining performance already done by Bouten in 2012 (P.C. Jos Bouten), but of attaining better performance by combining two types of forensic voice comparison systems.

Combining two systems

Let us look at different examples of automatic voice comparison systems:

System **A** has a very low error rate, but this error rate is only known to apply to very specific recordings, let's say long telephone-quality, single-channel recordings.

System **B** has a higher error rate, but this error rate is known to apply to a much wider variety of recordings, let's say recordings that may be fairly short, noisy, and include channel mismatch between suspect and question files.

Following van Leeuwen & Brümmer (2007), we can assess the performance of sys tems **A** and **B** not in terms of their error rates, but in terms of actual information extracted, which we express in "bits". The interpretation of these bits is related to the common C_{llr} error measure which is the average information loss of a system. An average of 0 bits means a C_{llr} of 1 and vice versa. If we have a same-spea ker-com parison and the system outputs a likelihood ratio (LR) of infinity, one bit is extracted (i.e. all the information available); if it outputs a LR of zero, zero bits are extracted (i.e. no information at all). For different-speaker-comparisons, the opposite is true.

Let's say System **A** only allows us to handle 10 cases a year, and for each it extracts 0.5 bits. This gives us a total extraction of 5 bits per year. Let's say System **B** only extracts 0.4 bits for each case, but it handles not just 10 but 25 cases a year. We obtain an average yearly total of 4 + 6 = 10 bits. What we can do now is set up a System **C** that combines Systems **A** and **B**: 10 cases will be extracted by its component **A**, 15 by its component **B**, and the total number of bits extracted yearly will be 5 + 6 = 11 bits. These examples are in line with real-life experience, as we will show using two current state-of-the-art voice comparison systems to simulate System **A** and a p-value approach which calculates scores without modelling intra-speaker variability (Solewicz et al. 2013) to simulate System **B**.

Discussion and Conclusion

While System **A**'s average performance remains better than **C**'s (and even more so of **B**'s), System **C**'s actual, total performance is better, since it handles two and half times as many cases as System **A** and is just as good as **A** in those cases that both can handle. However, depending on the field of interest, the expert might accept information loss while processing a desired number of cases or vice versa.

What we would therefore like to suggest is using simpler voice comparison algorithms, such as the one described in Solewicz et al. (2013), which produces scores without modelling intra-speaker variability. Such algorithms may at first seem like a step backwards when compared to state-of-the-art systems, but they could constitute a ma jor improvement to current practices in forensic speaker comparison, at least when com bined with these latter systems. For cases that meet certain, well-defined conditions, the expert will be able to extract maximum information; for cases that don't, the expert will be able to extract less information, which is better than none at all.

References

Przybocki, Mark A., Martin, Alvin F. and Le, Audrey N. (2007): NIST Speaker Recognition Evaluations Utilizing the Mixer Corpora – 2004, 2005, 2006. IEEE Transactions on Audio, Speech and Language Processing, **15**, 1951-1959.

Solewicz, Yosef A., Jardine, Gaëlle, Becker, Timo and Gfrörer, Stefan (2013): Estimated Intra-Speaker Variability Boundaries in Forensic Speaker Recognition Casework. Proceedings of Biometric Technologies in Forensic Science (BTFS) 2013, Nijmegen.

Van Leeuwen, David A. and Brümmer, Niko (2007): An Introduction to Application-Independent Evaluation of Speaker Recognition Systems. In Christian Müller (Ed.), *Speaker Classification I*, 330-353. Berlin. Springer.