Confidence Bounds Curves as a Tool for Evaluation of Automatic Speaker Recognition Results Uncertainty

Sergey Koval, Alexandra Lokhanova
Speech Technology Center, St. Petersburg, Russia
koval@speechpro.com,lohanova@speechpro.com

Abstract

In some applications of speaker recognition, for example in the forensic area or in the access control systems, an important task is to estimate some absolute measure of identity of the speakers. Automatic speaker recognition methods in this case seem to be the fastest and the simplest speaker identification tool [1-2]. However, up to now the applicability and reliability evaluation of automatic speaker recognition systems (ASRS) for single cases, e.g. in forensic area, is widely disputable [3-7]. Output results of state-of-the-art ASRS are based on statistical data analysis. Their applicability for individual comparisons is theoretically and practically rather complicated task. In this paper we address the issue of more detailed analysis of training data statistical structure and more careful decision making for ASRS in the context of one-to-one speech recordings comparisons using confidence bounds curve idea and bootstrap calculating technique.

1. Introduction

Up-to-date automatic speaker recognition systems (ASRS) are based on data-driven approaches of building and using speaker models. This is due to the fact that the factors affecting speaker characteristics are too numerous and complex in their influential structure for an exact definition in rules and constructive models [8-11]. Even if the training speech dataset is large enough, the learning outcomes of a particular ASRS can only say that if you repeat similar studies on this system, then the results would be similar to the results of the statistical learning of the system performed previously. This is acceptable and useful in practice if the ASRS is used, for example, to search for the target speaker in the list of a large number of suspects. On average, the system will correctly rank the speakers by their similarity with the target. However, in case the absolute results of a single comparisons are important, for example in forensic practice, usage of a point statistical estimation is insufficiently.

In theory and practice of identification analysis the international scientific society suggests solving such problems with a standard approach of uncertainty parameters in measurement [12]. In compliance with adopted definitions, uncertainty in measurement is “parameter, associated with the result of a measurement that characterizes the dispersion of the values that could reasonably be attributed to the measurand”. In this article we suggest to use the value of corresponding boundary of one-sided confidence interval (OCI) as such a parameter, as well as, relying on OCI, to determine curves of confidence boundaries (CB) of identification decision probability.

As adopted in multiple publications, we use Bayesian approach to interpret speaker recognition results [5,13-16]. For classifier scores received with ASRS we evaluate the posterior probabilities of identity of the speakers. This is the useful similarity measure for single comparisons because it has the clear interpretation as a degree of belief. To construct a confidence bound curve for posterior probability we use a nonparametric bootstrap method, which does not demand some presumptions about the shape of the sampling distribution. Bootstrap modeling is widely applied in ASRS performance evaluation to estimate the uncertainty of detection cost function [10], ROC curves [17] and FR-FA curves [18]. We show the importance of using confidence bound curves for the approaches which make single comparisons and need to estimate an absolute values of similarity.

2. Method

2.1. Posterior probability estimation

The task is to compare two audio records for the presence of speech belonging to the same speaker or different speakers. For the interpretation of a classifier score achieved by the ASRS for a

single comparison an expert need to have a set of "imposter" (different speakers) and "target" (same speakers) comparison results calculated with the same ASRS on some population. This population is usually chosen by the expert so as to take into consideration properties of the objective compared phonograms [1-2].Then the between-speaker and within-speaker variability distributions are estimated on achieved training set. According to this distributions the expert estimates some similarity measure between source phonograms. This estimation is the final result of ASRS work, applicable according to some authors in one or another form to court trials [1,15,19,20].

The Bayesian approach for speaker recognition implies the next two hypotheses:

H_0 - the suspected speaker is the source of voice on the questioned recording,

H_1 - the speaker at the questioned recording is not the suspected speaker.

Let us denote x as a classifier score achieved by the automatic method during the comparison of two recordings. For the treatment of measurement results in a probabilistic sense, it is convenient to use the posterior probability $P(H_0|x)$, i.e. probability of validity of the hypothesis about the identity of speakers under the supervision of a similarity measure x, or the likelihood ratio $LR = \frac{P(x|H_0)}{P(x|H_1)}$.

This two values are connected with the Bayes equation:

$$LR = \frac{P(H_0 \mid x) \cdot P(H_0)}{P(H_1 \mid x) \cdot P(H_1)}, \quad (1)$$

where $P(H_0)$ and $P(H_1)$ - prior probabilities of hypotheses, $P(x|H_0)$ and $P(x|H_1)$- likelihood of hypotheses. Posterior probability of the hypothesis, that the recordings contain voices of different speakers, is respectively equal to:

$$P(H_1|x) = 1 - P(H_0|x) \quad (2)$$

We propose to use the posterior probability $P(H_0|x)$ because it has the clear probabilistic sense and, moreover, it is useful for some post-processing, for example to make a fusion of results of different ASRS.

For the assessment of the posterior probability $P(H_0|x)$ we use the Platt calibration [21]. According to this method $P(H_0|x)$, can be represented as a sigmoid function of the score x:

$$P(H_0 \mid x) = \frac{1}{1 + e^{Ax+B}}, \quad (3)$$

where A, B – calibration parameters. The parameters are fitted using maximum likelihood estimation from a training set consisting of "imposter" and "target" comparison results.

2.2. Uncertainty estimation for the posterior probability

As a rule, the interpretation of the results of automatic identification techniques is limited to point measurements of the posterior probability or any other measure of similarity of speakers in the phonograms. However, in our opinion, this problem requires a more precise analysis of the situation. In order to give the more accurate interpretation of the identification results obtained by automatic methods, we refer to the concept of measurement uncertainty of posterior probability. To this end, we describe a procedure of constructing a one-sided confidence interval (OCI) - an interval that has a pre-specified high probability of including the true value of the parameter.

From statistical point of view the estimation of the true posterior probability $P(H_0 \mid x)$ is a random variable. To calculate CI of this estimation we can evaluate the parameters of its sample distribution. There are a number of methods to build CI, presuming Gauss shape of the sampling distribution, such as the Wilson interval, the Wald interval, etc. But for the speaker recognition task such a proposal is ungrounded.

The bootstrapping idea allows one to approximate sampling distribution based on only one initial sample. Let ‘s say an estimation of posterior probability $\hat{P}_n = \hat{P}_n(H_0 \mid x)$ is calculated for a starting sample $(X_1,...,X_n)$ using aforesaid algorithm. We need to construct for this estimation CI covering with given reliability the values of the evaluated parameter $P(H_0 \mid x)$. We construct a large number of repeated samples from one starting sample extracting with return its elements. Thus, given data is considered as the parent population to produce repeated samples. We create the set from B repeated samples $(X_1^*,...,X_n^*)$ and calculate for them correspondent values $\hat{P}_n^*(b), b = 1,..,B$. Based on estimates for bootstrap-samples we calculate the bootstrap-distribution – an analog of sample distribution for usual samples. Let cumulative distribution function of $\hat{P}_n^*$ is $G_*(p) = P\left\{\hat{P}_n^* \le p\right\}$. The bootstrap percentiles method gives quantiles of significance level α and $1-\alpha$ $G_*^{-1}(\alpha) = \inf\left\{x : G_*(x) \ge \alpha\right\}$ and $G_*^{-1}(1-\alpha)$ as lower and upper bounds $1-2\alpha$ -percentile CI for estimate $\hat{P}_n = \hat{P}_n(H_0 \mid x)$ [22]. We consider one-sided confidence interval (OCI). For the upper bound $\hat{P}_n^{upper} = G_*^{-1}(1-\alpha)$ the upper interval is:

$$P(-\infty < P(H_0|x) \le \hat{P}_n^{upper}) = \alpha, \tag{4}$$

and for the lower bound $\hat{P}_n^{lower} = G_*^{-1}(\alpha)$ the lower interval is:

$$P(\hat{P}_n^{lower} \le P(H_0|x) < \infty) = \alpha. \tag{5}$$

For comparison with a high similarity degree we use the value of lower CI and for comparison with law similarity degree we use the value of upper CI.

The bootstrap method proposes that elements in original sample are independent. But in our case the sample elements are dependent when we compare records of the same speakers. For this problem solving we use Subset bootstrap [18]. The subset bootstrap proposes to fragment the all comparisons set into independent comparisons subsets for every speaker's pair. The repeated sample is formed from subsamples for each of these subsets.

3. Experimental results

We demonstrate some experimental results to investigate the influence of measurement uncertainty on the interpretation of ASRS results. We use the NIST 2008 speaker recognition evaluation dataset and The Russian Switched Telephone Network corpus (RuSTeN) [23]. As ASRS we use GMM-SVM-based system, developed by Speech Technology Center, Ltd (STC) for NIST SRE 2010 [11].

3.1. Confidence interval calculation for the posterior probability

We use the sample from NIST 2008 SRE dataset, where there are microphone recordings of interview. To make this data more uniform we use only the male voices. According to (2.1) using Platt calibration we estimate $P(H_0 \mid x)$, i.e. the posterior probability that for given classifier score x speakers are the same. Then, according to (2.2) we use the bootstrap approach to estimate upper and lower CI bounds in every point x for significance level $\alpha = 0.80, 0.95$ and 0.99. It is important for voices comparisons with high similarity degree to get lower bound of $P(H_0 \mid x)$, that $P(H_0 \mid x) \ge \hat{P}_n^{lower}$ with some confidence level α. And vice versa, for voices comparison with low similarity degree to get upper bound: $P(H_0 \mid x) \ge \hat{P}_n^{upper}$ with some α. We find out the area of x values, where with probability α it is possible to assert that the speaker is rather “friend” than “foe. i.e. when lower bound of $\hat{P}_n^{lower} > 0.5$. Similarly one can find out the area of x, where with probability α it is possible to assert that the speaker is rather “foe” than “friend. i.e. when upper bound $\hat{P}_n^{upper} < 0.5$. The intermediate area corresponds to an uncertain situation. Figure 1 shows

dependence $P(H_0 \mid x)$ from classifier score x for sample from NIST 2008 SRE dataset, as well confidence bound (CB) curves for $P(H_0 \mid x)$, which are limits of upper-CI for left area and lower-CI for right. We can note that CB for $P(H_0 \mid x)$ gives a significantly less strong decision for speakers similarity degree, than standard estimate $P(H_0 \mid x)$, but reliability degree of the identity decision now may be exactly characterized by significance level. The middle area with horizontal CB curve is in this case an uncertainty zone, which appears due to taking into account the measurement uncertainty.

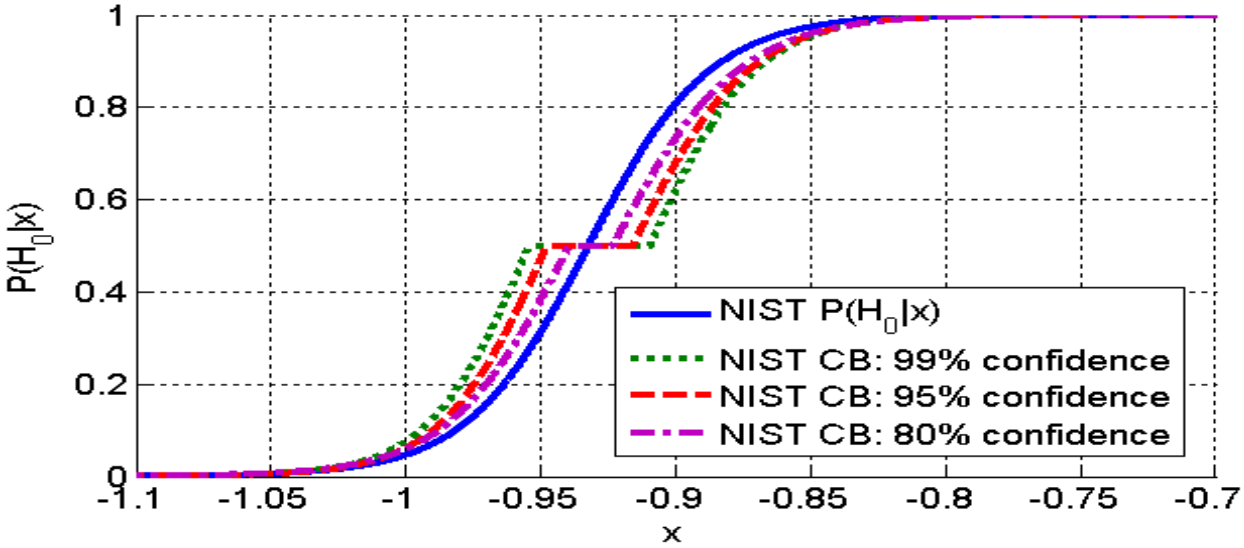

Fig. 1: *Probability* $P(H_0 \mid x)$ *versus x-score curve and confidence bound (CB) curves for* $P(H_0 \mid x)$ *calculated for sample from NIST 2008 SRE for confidence levels 80%, 95% and 99%.*

The proposed approach gives the user for every individual case their own values of minimal match probability and maximal mismatch probability for every confidence level. Except that for the given ASRS the new feature appears: between voices similarity measure area, where the identity decision for given confidence level is in principle uncertain.

3.2. Confidence interval dependence from training speech dataset

We investigate how we will change the CB for $P(H_0 \mid x)$ in dependence from speech material, used for evaluation. We used the samples from the NIST2008 SRE dataset and the speech database RuSTeN [23]. The one sample from the NIST2008 SRE dataset includes remote microphone recordings, the other sample includes telephone and RuSTeN - landline telephone recordings. This datasets have a different level of quality: the telephone sample from the NIST SRE dataset has the best quality and phonograms from the RuSTeN - the worst. The average signal-to-noise ratio (SNR) for this datasets is shown in Table 1.

Table 1. Average Signal-to-Noise Ratio (SNR) for training datasets

	NIST telephone	**NIST microphone**	**RuSTeN**
SNR	28 dB	18 dB	15 dB

Fig. 2 shows evaluation of posterior probability $P(H_0 \mid x)$ for different training datasets, and confidence bounds (CB) curve for $\alpha = 0.95$.

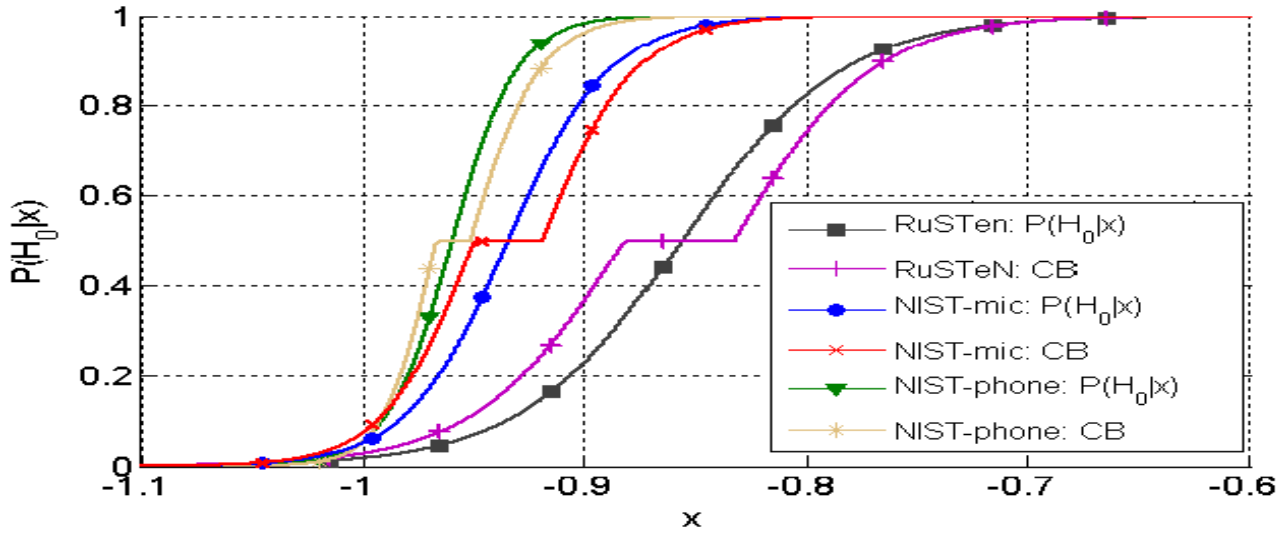

Fig. 2*: Probability $P(H_0|x)$ versus x-score curves and confidence bound (CB) curves for $P(H_0|x)$, calculated for samples from NIST 2008 SRE phone, microphone and RuSTeN base for α=0.95.*

First, we can say that $P(H_0 \mid x)$ values depends dramatically on the training dataset. Second, the worse the quality of training records is, the larger the CI and the uncertainty zone (horizontal curve area) are. In case we use the RuSTeN database as a training set we have around 2-4 times more uncertain result, then for training dataset with good sound quality. So the end ASRS user is responsible for the adequate choice of the ASRS training speech data base for a given case, and ASRS developers are responsible for accordance of needed possibilities.

4. Conclusion

We showed the practical way of confidence bounds curves usage for evaluating ASRS results uncertainty. OCI were estimated with the bootstrap method. We demonstrated that this framework may be effectively used for one-to-one speech files comparison, in particular, in forensic applications of ASRS. This approach to evaluation and interpretation of ASRS results supplements the usage of traditional ROC, DET, TIPPET curves and allows for greater reliability to be applied to statistical conclusions for single cases.

5. References

1. *Drygajlo A.* Forensic automatic speaker recognition. IEEE Signal processing Magazine, 24(2):132-135, 2007.
2. *Meuwly D., El-Maliki M. and Drygajlo A.* Forensic Speaker Recognition Using Gaussian Mixture Models and Bayesian Framework. COST 250 Workshop on Speaker Recognition by Man and Machine, Directions for forensic applications, 52-55, Ankara, Turkey, Apr 1998.
3. Interspeech 2008 special session “Forensic Speaker Recognition Traditional and Automatic Approaches“. Online: http://interspeech2008.forensic-voice-comparison.net, accessed on 22-26 Sep 2008.
4. *Bonastre J.-F., Bimbot F., Boe L.-J., Campbell J. P., Reynolds D. A. and Magrin-Chagnolleau I.* Person Authentication by Voice: A Need for Caution. In Proc. Eurospeech in Geneva, Switzerland, ISCA, 33–36, Sep 2003.
5. *Rose P.* Technical forensic speaker recognition: Evaluation, types and testing of evidence. Computer Speech and Language 20, 2-3: 159-191, 2006.
6. *Nolan F.* Speaker identification evidence: its forms, limitations and roles. In Law and Language: Prospect and Retrospect, Levi, 2001.
7. *Eriksson A.* Tutorial on forensic speech science. Part 1. Forensic phonetics. In Interspeech 2005 - Eurospeech. Proceedings of the 9th European Conference on Speech Communication and Technology, Lisbon, Portugal, 2005.
8. *Martin A.F., Greenberg C.S.* The NIST 2010 Speaker Recognition Evaluation. INTERSPEECH 2010, 2726-2729, Makuhari, Chiba, Japan, 26-30 Sep 2010.

9. *Ferrer L., Graciarena M., Kajarekar S., Scheffer N., Shriberg E., Stolcke A.* The SRI NIST SRE10 Speaker Verification System. NIST Speaker Recognition Evaluation Workshop, Brno, Czech Republic, 24 Jun 2010.
10. *Wu J.; Martin A. F.; Greenberg C. S.; Kacker R. N.* Measurement Uncertainties in Speaker Recognition Evaluation. NIST Publication, 1-16, 15 Sep 2010.
11. *Belykh I.N., Kapustin A.I., Kozlov A.V., Lohanova A.I., Matveev Yu.N., Pekhovsky T.S., Simonchik K.K., Shulipa A.K.* The speaker identification system for the NIST SRE 2010. Will appear in "Informatics and its Applications. 5(4), 2011.
12. Guide to the Expression of Uncertainty in Measurement, International Organization for Standardization (ISO), Geneva, 1993.
13. *Evett I., Buckleton J.* Some aspects of the Bayesian approach for evidence evaluation. Journal of Forensic Science Society, 29:317-324, 1989.
14. *Meuwly D., Drygajlo A.* Forensic speaker recognition based on a Bayesian framework and Gaussian mixture modeling GMM. In Proc. Odyssey01, 145-150, 2001.
15. *Gonzalez-Rodriguez J, Garcia-Romero D., Garcia-Gomaran M., Ramos-Castro D., Ortega-Garcia J.* Robust likelihood ratio estimation in Bayesian forensic speaker recognition. In Proc. Eurospeech, 693-696, 2003.
16. *Campbell W. M., Reynolds J. P., Campbell D. A., Brady K. J.* Estimating and evaluating confidence for forensic speaker recognition. In Proc. ICASSP2005, Philadelphia, PA, 2005.
17. *Poh N., Martin A., Bengio S.* Performance Generalization in Biometric Authentication Using Joint User-Specific and Sample Bootstraps. IEEE Trans. on PAMI, 29(3):492-498, 2007.
18. *Bolle R.M., Ratha N.K. and Pankanti S.* Error Analysis of Pattern Recognition Systems: the Subsets Bootstrap. Computer Vision and Image Understanding, 93(1):1–33, 2004.
19. *Fesenko A.V.* [Ed]. Speakers identification by Russian speech recordings using the automated system "Dialect". Moscow: Military unit 34435, in Russian, 1996.
20. *Drygajlo A.* Statistical Evaluation of Biometric Evidence in Forensic Automatic Speaker Recognition. 3rd International Workshop on Computational Forensics (IWCF 2009), The Hague, The Netherlands, 2009.
21. *Platt J.* Probabilistic outputs for Support Vector Machines and comparisons to regularized likelihood methods. Advances in Large Margin Classiers, Cambridge: MIT Press, 1999.
22. *Efron B., Tibshirani R.J.* An Introduction to the Bootstrap. Chapman & Hall, New York, 1993.
23. LDC 2006S34 ISBN 1-58563-388-7, www.ldc.upenn.edu.