Background Specificity in Forensic Voice Comparison and Its Relation to the Bayesian Prior Probability

Michael Wagner[1], Yuko Kinoshita[2]

[1]Faculty of ESTeM, University of Canberra
[1]Research School of Computer Science, The Australian National University
[1]Quality and Usability Lab, Technical University of Berlin
[2]College of Arts and Social Science, The Australian National University
michael.wagner@canberra.edu.au; yuko.kinoshita@anu.edu.au

Abstract

This study investigates the effect of background data specificity on likelihood ratio and prior odds, and consequently on the posterior odds outcome. It is motivated by discussions on the correct choice of speaker recognition background, particularly in forensic voice comparison. We performed strictly controlled experiments with the ANDOSL database where background specificity is the sole independent variable. Results show that target and non-target scores are better separated with less specific background, but that in turn priors must be adjusted down. Because the risk of class recognition instead of individual recognition increases with lower background specificity, we suggest that the prior probability in the Bayes formula is factorised into one part that remains in the domain of the trier of fact – as is conventional – and another part that is related to the specificity of the assumed or agreed background.
Index Terms: Forensic voice comparison, Bayesian method, forensic prior probability, background specificity.

1. Introduction

The Bayesian approach used in forensic voice comparison (FVC) is similar in principle to that of non-forensic speaker authentication tasks. However, the interpretation of Bayesian likelihood ratios (LRs) is quite different in the FVC context.

In non-forensic speaker authentication, the Universal Background Model (UBM) has long been the standard method [1, 2, 3, 4]; contemporary methods, such as joint factor analysis, iVectors etc. are also implicitly based on the chosen UBM. Systems such as those prominent in recent NIST speaker recognition evaluations [5], generally use very large UBMs that represent speaker characteristics across dialects, accents and even languages spoken in multicultural societies, except that by common consensus, they are usually restricted to speakers of the same sex as that of the unknown speaker. To compensate for external factors, such as environmental noise and channel characteristics, speaker recognition scores are normalised with a cohort of speakers with similar attributes.

In FVC, the situation is somewhat more complicated. Within the context of the widely accepted Bayesian paradigm [6, 7], the forensic speech scientist estimates the likelihood ratio (LR) between the two likelihoods: 1) for the crime-scene recording to be consistent with the speaker model of the suspect (numerator of the LR) and 2) for the same recording to be consistent with the multi-speaker model of a background population (denominator of the LR). Although it is rarely performed explicitly in reaching the final decision, the trier of fact is required to combine the LR obtained from FVC with the prior odds $P(H_{so})/P(H_{do})$: the prior probability of the same-origin hypothesis (H_{so}–offender and suspect are the same person) versus the prior probability of the different-origin hypothesis (H_{do}–offender and suspect are different persons).

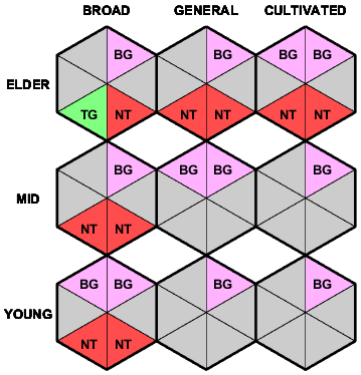

Figure 1. *Partitioning of the non-accented male speakers of ANDOSL into 9 subgroups, each of 6 speakers, and speaker partitioning into target (TG), non-target (NT) and background (BG) speakers.*

The determination of the prior odds is usually considered the domain of the court, as the forensic scientist does not have access to information on the case other than the voice recordings. Combining those prior odds with the forensic scientist's LR $p(X|H_{so})/p(X|H_{do})$ yields the posterior odds $P(H_{so}|X)/P(H_{do}|X)$ for the same-origin hypothesis versus the different-origin hypothesis according to the Bayes Rule [7]

$$\frac{P(H_{so}|X)}{P(H_{do}|X)} = \frac{P(H_{so})}{P(H_{do})}\frac{p(X|H_{so})}{p(X|H_{do})}. \quad (1)$$

In FVC, H_{do} plays a pivotal role in selecting the background population, and this has been the subject of much debate among forensic scientists. Some have argued that the background population should be tailored to the characteristics of the offender, the suspect or both [8]. It has also been argued that the background population should be based on those characteristics of the offender's voice that both prosecution and defence agree upon [6] or that it should represent a set of speakers sufficiently similar to the offender's voice that an investigating police officer would bother submitting voice samples for examination by a forensic scientist at all [9].

In forensic casework, the different-origin hypothesis defines the subpopulation to which the offender apparently belongs and to which any suspects should also belong. Those attributes of the subpopulation that can affect speech acoustics (e.g. language variety, gender, age range) will set the selection criteria for the background population data for the case. As should be clear from Eq. (1), the selection of the background affects the forensic scientist's LR through $p(X|H_{do})$ as well as the trier of fact's assignment of prior odds through $P(H_{do})$.

A commonly used imaginary forensic examination illustrates the two effects: Assuming a criminal investigation on an island of 100 inhabitants, if nothing other than being a member of the island population is known about the offender, the prior odds $P(H_{so})/P(H_{do})$ would be 1/99, and the forensic scientist should build a background model from a representative sample of the entire island population. If however, in addition, the offender were known to be a member of the female half of the population, the prior odds would increase to 1/49, and the forensic scientist would build the background model only from the female subpopulation. Any additional knowledge about the offender would further raise the prior odds and, at the same time, likely diminish the acoustic variance of the background.

While the characteristics of the background population such as gender or dialect should be consistent with an agreed H_{do}, in practice the forensic scientist's choices are often constrained by data availability, time and resources. Sometimes there is not even a clear reference to an agreed H_{do}. This is problematic for the validity of the Bayesian estimation of the posterior odds unless the scientist informs the trier of fact of this relationship between H_{do} and choice of background data on one hand, and LR, prior and posterior odds on the other.

In the remainder of this paper, we thus examine how the specificity of the different-origin hypothesis and the corresponding selection of the background population affect the outcome of forensic voice comparison: firstly through $p(X|H_{do})$ and the resulting LR and secondly through $P(H_{do})$ and its effect on the prior odds estimated by the trier of fact.

2. Experiment

2.1. Data

The Australian National Database for Spoken Language (ANDOSL) [10] comprises Australian English speech data from a range of speakers, varying in age, sex, and their variety of Australian English. For this study, we utilise only the read-sentence data by the ANDOSL male native speakers. Within that population, we have a 3×3×6 partitioning into 3 age groups, elder, mid, young, the 3 sociolect groups of Australian English, broad, general, cultivated, on the basis of the tag provided in ANDOSL [11], and 6 speakers in each group, as is illustrated in Fig. 1.

Each of the 9 subpopulations is shown as a hexagon, and the 6 speakers of each subpopulation are shown as colour-coded triangles. The single target speaker is shown as the green triangle, tagged *TG*. The non-target speakers are shown as red triangles, tagged *NT*, 5 in the first row for the *elder* subpopulation and 5 in the first column for the *broad* subpopulation. The 12 background speakers, tagged *BG*, are shown as magenta triangles. This design ensures that there is no overlap between target, non-target and background speakers, hence avoiding a potential statistical bias.

Each of the 54 male speakers read 200 sentences that were designed to cover the entire acoustic-phonetic space of Australian English. Of those, 180 were used solely for training background models. Using 180 sentences for UBM training, we assume that the background models cover the acoustic-phonetic space of the background population sufficiently. Of the remaining 20 sentences, 10 were used solely for maximum-a-posteriori (MAP) adaptation of the target-speaker models, and 10 were used solely for the target and non-target testing. We consider using 10 sentences for GMM adaptation forensically realistic, given the typical constraints of FVC casework, where suspects are often uncommunicative during police interviews and provide precious little material for the adaptation of the target-speaker GMM.

Recordings of the 3×3×6×200=10,800 sentences are stored as wav files, sampled at 20,000 samples/s and 16 bits/sample. 12 mel-frequency cepstral coefficients (MFCC) and log energy were determined for 20ms windows shifted in 10ms steps. Derivative coefficients were discarded as the amount of data was insufficient for training higher-dimensional models. Using a simple absolute energy threshold, low-energy frames were eliminated and about 61% of the frames retained, yielding on average 313 feature vectors per sentence for the analysis.

2.2. Experimental design

The read-speech data in ANDOSL were produced under highly controlled conditions, and each speaker was recorded in a single session. ANDOSL is thus generally regarded as an inadequate database for FVC experiments. However, our experimental design turns this limitation into an advantage. The single-session nature of ANDOSL and the read-speech material enabled perfect control over the independent variables of our design. Being a single session recording eliminates extraneous variation such as intersession and channel variation as well as intra-speaker variation in health or emotion. Using speech material read from prepared texts, we exclude the variability in quantity and phonetic contents that is inevitable with spontaneous speech data. Therefore, we can reasonably interpret any effects on the output as being caused by the chosen background specificity–the independent variable of our design.

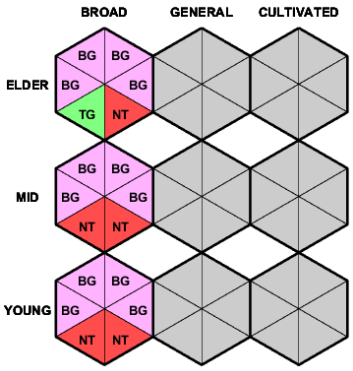

Figure 2. *Speaker verification in the broad subpopulation with specific background.*

We conducted altogether 4 experiments. In the first, the target speaker, the 5 non-target speakers and the 12 background speakers are all from the subpopulation of the *broad* sociolect speakers as shown in Fig. 2. In the second experiment, the target speaker, the 5 non-target speakers and the 12 background speakers are all from the *elder* subpopulation as shown in Fig. 3. Each experiment proceeded to build a UBM

from the background speakers, building a GMM for the target speaker, and determining LRs for target and non-target tests.

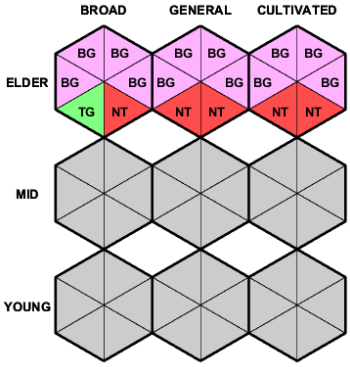

Figure 3. *Speaker verification in the elder subpopulation with specific background.*

We then repeated these 2 experiments altering the specificity of the UBM by building it from the background speakers drawn from the entire population of male speakers as shown in Fig. 1, while the other factors, i.e. target and non-target speakers and the sentence material, were kept identical. For the comparison of system performance, LLRs are usually normalised or calibrated for environmental or channel variation between recordings. However, the current study does not require such a step due to the strict control in experimental design.

Fig. 4 depicts our experimental design schematically. We constructed a UBM from the 180 designated sentences spoken by the 12 chosen background speakers. For the target speaker, we MAP-adapted this UBM to a target-speaker-specific GMM using the 10 designated adaptation sentences of the target speaker. Finally, 10 target trials were conducted with the designated test sentences of the target speaker, and 50 non-target trials were conducted with the same test sentences spoken by 5 non-target speakers. For each trial, we produced the sentence-mean log likelihoods and log likelihood-ratios with respect to the target GMM and the UBM.

For the *broad*-sociolect subpopulation, we conducted the following pair of experiments: firstly, we built a UBM from the 12 designated background speakers of the subpopulation as illustrated in Fig. 2. For the designated target speaker, we adapted the UBM to build the target-speaker GMM and conducted the target trials. Then we conducted the non-target trials with the designated 5 non-target speakers of the subpopulation. Both target and non-target trials consist of the designated 10 testing sentences for each target and non-target speaker. That experiment was repeated with the same target and non-target speakers, but with the 12 background speakers drawn from the full population as shown in Fig. 1 and the target-speaker GMMs adapted from that wider-background UBM.

An equivalent set of experiments was then conducted with the other subpopulation under investigation, *elder* speakers. Here, the background speakers were drawn from this subpopulation and from the full population as shown in Fig. 3 and again in Fig. 1. For both pairs of experiments, the independent variable is the specificity of the background: either specific to the subpopulation matching the test speaker characteristics or less specific by being a superset of that subpopulation–in our case the entire population of the male speakers in ANDOSL.

For each trial, we observe the log likelihoods (LL) $log\ p(X|H_{so})$ and $log\ p(X|H_{do})$ for each sentence X, each being the mean of the frame log likelihoods for the sentence. We also observe the resulting log likelihood-ratios $LLR = log\ p(X|H_{so}) - log\ p(X|H_{do})$. The statistics of the above LLs and LLRs are the dependent variables of the design, while background specificity is the independent variable.

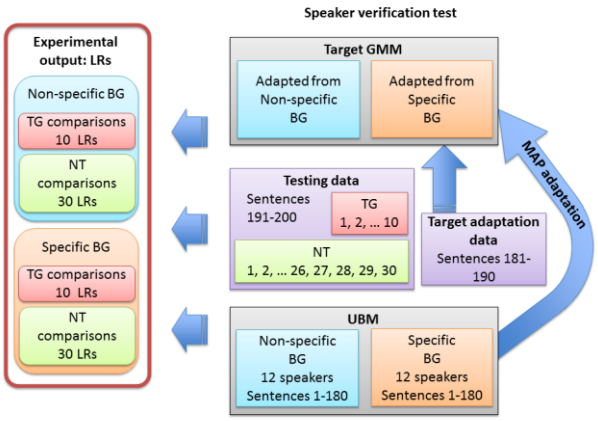

Figure 4. *Experimental speaker recognition system showing the parallel evaluations of non-specific background (blue boxes) and specific background (orange boxes)*

3. Results and discussion

Table 1 presents the mean LLRs for the target and non-target tests for the 2 sets of experiments and their mean differences ΔLLR as well as their log-likelihood-ratio costs C_{llr} [12]. The last 2 rows combine the 2 specific-background and the 2 non-specific-background experiments.

Table 1. *Target and non-target LLRs and their differences.*

Target/UBM	TG LLR	NT LLR	ΔLLR	Cllr
Broad/Broad	2.664	-0.174	2.837	0.496
Broad/Non-specific	2.921	-0.174	3.096	0.491
Elder/Elder	2.682	-0.409	3.091	0.420
Elder/Non-specific	2.921	-0.311	3.232	0.438
Mean Specific	2.673	-0.291	2.964	0.458
Mean Non-specific	2.921	-0.243	3.164	0.464

The results show that the target and non-target scores are separated better for the non-specific background than for the specific background. Fig. 5 shows in addition the distribution of the numerator LLs (LLG) and the denominator LLs (LLU) in Eq. (1). The curves are based on the means and variances of the LLs and a normality assumption for the distributions.

The 2 largely overlapping narrow (green) Gaussians at the right of Fig. 5a show the distributions of the numerator LLs of the *broad* target trials against the specific UBM (dashed line) and against the non-specific UBM (full line). The specificity of the UBM seems to affect neither the mean nor the variance of those LLs. The other 2 narrow (black) Gaussians near the centre of Fig. 5a show the distributions of the denominator LLs of the *broad* target trials against the specific and non-specific UBMs. Those distributions show that the non-specific UBM produces distinctly smaller denominator LLs than the specific UBM. Since the numerator LLs are distributed almost identically, it follows that the non-specific UBM produces higher LLRs than the specific UBM.

Fig. 5b shows the corresponding 4 curves for numerator and denominator LLs against specific and non-specific UBMs for the *elder* subpopulation with the same trends as found for Fig. 5a. Each of the 2 figures also shows 4 wide Gaussians for

the non-target trials with the respective specific numerator (green) and denominator (black) LLs closely overlapping and the specific UBM (dashed) yielding a slightly larger mean LL than the non-specific UBM (full) as could be expected.

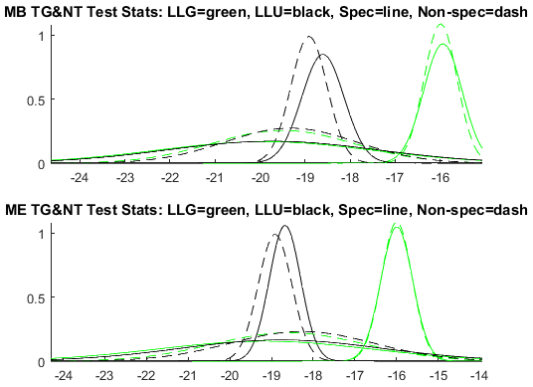

Figure 5. *LL statistics of target trials (narrow) and non-target trials (wide) against target GMM (green) and UBM (black): (a) Broad subset; (b) Elder subset (Note the almost complete overlap between the 4 pairs of wide curves in the 2 figures)*

In both cases, the background specificity affected non-target comparisons less. Our results show a larger distance between target and non-target scores for the non-specific UBM, which seems to be due to the larger denominator LLs of the non-target scores for the non-specific UBM. Table 1 shows no significant differences of C_{llr} between the specific and the non-specific background in either experiment.

From this rather small and therefore limited study, it appears that the specificity of the background is not of major consequence in FVC or even that a less specific background may be preferable for reasons of the slightly larger distance between target and non-target scores found here. However, this interpretation must be weighed against 2 other factors: Firstly, a less specific UBM has the tendency to turn the speaker recognition problem into a recognition of the subpopulation. In other words, there is a danger of recognising, for example, the language variety of a speaker instead of recognising the individual. And secondly, a less specific UBM implies a proportionately smaller prior probability for the defence hypothesis and correspondingly a less conclusive outcome in terms of the posterior odds of the analysis.

For example: an offender is only known to be a member of a population of 8 million Australian adult males. Non-acoustic evidence such as height, eye and hair colour place him in 10% of that population. Using Row 2 of Table 1, the forensic scientist reports a likelihood ratio of $e^{3.096} = 22.109$ against that background population with corresponding posterior odds of $1/7{,}999{,}999 \times 10\% \times 22.109 = 0.276 \times 10^{-6}$.

However, if on the acoustic evidence the forensic scientist determines that the offender is a member of the *broad* accent group of 2 million males and, according to Row 1 of Table 1, reports a likelihood ratio of $e^{2.837} = 17.064$ against that more specific background population, the corresponding posterior odds are $1/1{,}999{,}999 \times 10\% \times 17.064 = 0.853 \times 10^{-6}$, a value about 3 times larger than with non-specific background.

This example illustrates the case for factoring the prior odds into one part that represents the non-acoustic evidence (10% in our example) and another part that represents the size of the background population used by the forensic scientist.

4. Conclusions

A small-scale preliminary study was conducted to investigate how the specificity of the background population affects FVC, using age and sociolect specific subpopulations in the ANDOSL database. LLRs as well as the constituting numerator and denominator LLs were determined dependent on background specificity. Our small-scale experiments show that the denominator LLs for the target speaker were smaller for less specific UBMs and hence those LLRs were larger and the target-non-target separation was larger for less specific UBMs. However, a less specific background bears the risk of inadvertently performing class recognition instead of individual recognition. Further experiments with larger datasets and varying degree of specificity should be conducted.

Also, perhaps more importantly, we must be mindful that the choice of background population directly affects the determination of the prior odds and thus the interpretation of the forensic voice comparison by the trier of fact. In the example presented in this study, assuming equal distribution of the subpopulations, the choice of the full male population of ANDOSL for the UBM would imply prior odds 3 times smaller than the choice of the specific subpopulations.

It is therefore most important for the forensic scientist to report as precisely as possible the characteristics of the background database and its implications for the determination of the LRs and posterior odds of the forensic voice comparison.

5. References

[1] A.E. Rosenberg, J. DeLong, C.H. Lee, B.H. Juang, F.K. Soong, "The use of cohort normalized scores for speaker verification", *Proc. Int. Conf. on Spoken Language Processing*, 599-602, 1992.

[2] J.B. Millar, F. Chen, M. Wagner, X. Zhu, "The efficacy of cohort normalisation in a speaker verification task under different types of speech signal variance", *Proc. Austr. Int. Conf. on Speech Science and Technology*, 850-855, 1994.

[3] S. Furui, "Recent advances in speaker recognition", *Proc. 1st Int. Conf. on audio- and video-based biometric person authentication*, 237-252, 1997.

[4] D.A. Reynolds, T.F. Quatieri, R.B. Dunn, "Speaker verification using adapted Gaussian mixture models", *Digital Signal Processing*, 10, 19-41, 2000.

[5] G.R. Doddington, M.A. Przybocki, A.F. Martin, D.A. Reynolds, "NIST speaker recognition evaluation - overview, methodology, systems, results, perspective", *Speech Communication*, 31, 225-254, 2000.

[6] P. Rose, *Forensic speaker identification*, London: Taylor and Francis, 2002.

[7] G.S. Morrison, "Forensic voice comparison" in I. Freckelton and H. Selby [eds], *Expert Evidence*, Ch. 99, Sydney: Thomson Reuters, 2010.

[8] N. Brümmer, E. de Villiers, "What is the 'relevant population' in Bayesian forensic inference?", downloaded on 30 March 2016 from http://arxiv.org/pdf/1403.6008v1.pdf, 2014.

[9] G.S. Morrison, F. Ochoa, T. Thiruvaran, "Database selection for forensic voice comparison", *Proc. Odyssey 2012*, 62-77, 2012.

[10] J. Vonwiller, I. Rogers, C. Cleirigh, and W. Lewis, "Speaker and material selection for the Australian national database of spoken language", *Journal of Quantitative Linguistics*, 2, 177-211, 1995.

[11] J. Harrington, F. Cox, and Z. Evans, "An acoustic phonetic study of broad, general, and cultivated Australian English vowels," *Australian Journal of Linguistics*, 17:2, pp. 155-184, 1997.

[12] N. Brümmer, J. du Preez, 2006. "Application-independent evaluation of speaker detection," *Computer Speech & Language*, 20, 230–275.