A test of the effectiveness of speaker verification for differentiating between identical twins

Aladdin Ariyaeeinia[1], Christopher Morrison[2], Amit Malegaonkar[1], Sue Black[2]

[1] University of Hertfordshire, College Lane, Hatfield, AL10 9AB, UK
[2] University of Dundee, Dundee, DD1 4EH, UK
a.m.ariyaeeinia@herts.ac.uk, morrisonc@medimmune.com,
amalegaonkar@trinityconvergence.com, s.m.black@dundee.ac.uk

ABSTRACT

This paper presents investigations into the ability of speaker verification technology to discriminate between identical twins. It is shown that whilst, in general, the genetic and non-genetic characteristics of voice are both of value to speaker verification capabilities, it is the latter which is highly beneficial in the separation of the speech of identical twins. It is further demonstrated that through the use of unconstrained cohort normalisation as a complementary means for the exploitation of such voice characteristics, the verification reliability can be considerably enhanced for both identical twins and unrelated speakers. Experiments were conducted using a bespoke clean-speech database consisting of utterances from forty nine identical twin pairs. The paper details the problem in speaker verification posed by identical twins, discusses the experimental investigations and provides an analysis of the results.

INTRODUCTION

Speaker verification (SV) is a principal subclass of speaker recognition (voice biometrics), defined as determining whether a speaker is who s(he) claims to be, based on a presented sample utterance. This has been the subject of extensive research in recent years [1-2]. In practical terms, the main goal for speaker verification is to minimise the overlap between the score distributions for a) the true speaker and b) the impostor, to reliably verify or reject a claimed identity using a preset threshold. An area of concern in this process, which has been the focus of attention over the past decade, is that of variation in speech

characteristics. Such variation can have different causes including ambient noise and uncharacteristic sounds generated by the speakers (e.g. lip smacks and mouth clicks). The resultant variation in speech can cause a mismatch between the presented utterance and the pre-stored voice pattern recording for the genuine speaker. Such mismatches have undesirable effects on the score distribution parameters for the true speakers and this can, in turn, lead to further overlapping of the score distributions for the true speaker and for the impostors who are targeting that particular speaker. In practical applications of automatic speaker verification, it is not normally possible to gather accurate information on the existence, level and nature of speech variation. In such cases, the most effective way to deal with this problem is score normalisation [2–6]. To date, a number of normalisation techniques have been developed, which are based on either the Bayesian approach or the standardisation of the score distributions.

An important issue in the field of automatic speaker verification (SV) is the potential challenge posed by identical (monozygotic) twins. The expectation of this challenge is due to the general concept that monozygotic twins should be highly similar in every respect including their voices. Although there have been some previous investigations into the effectiveness of automatic voice discrimination for such an application, these have been generally lacking in terms of the database used, the capability of the technology deployed, or both [7-10]. The aim of this study is to examine the capability of the current state-of-the-art speaker verification for discriminating between identical twins. Additionally, the study is based on using a database consisting of speech from a relatively large set of appropriately verified identical twins (i.e. 98 speakers).

When offspring are genetically identical, i.e. they have developed from the same fertilized zygote that has split, they are referred to as monozygotic twins. Dizygotic siblings are not genetically identical and arise from separate fertilization events through multiple oocyte release. For multiple births of more than two offspring there can be a combination of monozygotic and dizygotic individuals. The mechanisms which

give rise to multiple births vary depending upon the zygosity of the offspring. For dizygotic siblings, fertilisation takes place in the same way as it would for a singleton, with the notable difference that two or more oocytes are released from the ovaries at approximately the same time. As each female gamete is fertilised by separate male gametes, the resulting offspring are not identical and share only 25% of their genes - assuming paternal consistency [11]. Dizygotic siblings are not always of the same sex and are more commonly referred to as fraternal or non-identical siblings. The fetuses in such a multiple birth do not generally share any of their fetal membranes, each having their own placenta, amnion and chorion, although exceptions do occur [12].

Monozygotic siblings arise from the cleavage of a single fertilised egg and, being genetically identical, are generally referred to as identical siblings [11]. Although it is possible, in theory, for monozygotic siblings to develop as entirely separate embryos due to a very early division of a two-cell embryo, it is believed to be more common for identical siblings to develop from the separation of the inner cell mass at the pre-implantation blastocyst stage i.e. 4-6 days post-fertilisation resulting in a greater likelihood of shared fetal membranes [11, 13]. When an embryo splits after eight days, complete separation of the embryos is unlikely, resulting in conjoined or Siamese twins.

Such variation ensures that zygosity cannot be determined with any complete accuracy solely by documenting the sharing or otherwise of the placental support structures and therefore other non-invasive methods for determining whether siblings are monozygotic or dizygotic are required. Determining zygosity through DNA analysis is of course a more reliable option but this is an invasive and costly process and something that is unlikely to be permitted for ethical reasons in most studies of twins. A questionnaire known as the ‘peas in a pod’ or ‘PPQ’ has been shown to be 95% accurate in determining zygosity [14]. The PPQ firstly asks siblings to confirm their birth gender and then asks five questions relating to other people’s ability to distinguish between them when they were younger. A scoring system

is used to determine zygosity, with scores 0-3 indicating monozygosity and 8-10 dizygosity. However, the scores of each sibling must be in agreement for zygosity to be determined with any reliability. If scores are in disagreement or a score of 4-7 is recorded then zygosity is not obvious from physical appearance and is recorded as unknown [15].

The 'nature - nurture' argument associated with multiple births has always held a fascination with scientists from many disciplines and this has now been transferred to the biometrics arena. In a field where security systems are specifically designed to maximise individuality, monozygotic siblings offer an interesting paradox of being identical in the vast majority of their biometric characteristics, yet presenting as more than one individual.

In general, the study of identical twins is of interest to academics from a wide range of disciplines as it allows the exploration of the role that the genetic and environmental factors play on our development. In speaker verification, the challenge expected from monozygotic siblings is based on the anticipation that they should sound identical due to their physiological development and also the assumption that they have been exposed to identical environmental factors. However, as they age (or are separated) and experience greater independence, their voices are subjected to extraneous influences. This results in differences in the non-genetic characteristics of their voices, that can be of physiological as well as habitual nature. For instance, cigarette smoking can have a significant effect on the voice [16], and geographical separation may lead to dialect variation. Such differences have already been investigated and reported in a number of phonetic studies [17-18].

It should be noted that, in the present study, the voice genetic/non-genetic characteristics used for speaker discrimination are mainly of physiological nature (vocal tract) rather than habitual nature (e.g. dialects). This is due to the use of short-term speech features in the state-of-the-art speaker modelling and classification. Whilst the dissimilarities of this nature can be beneficial in automatically differentiating

between identical twins, their usefulness could be better exploited if there was the possibility for directly accessing such information. However, such data is encoded in speech and captured only implicitly in the speech features. An approach to this problem is thought to be through the use of UCN (unconstrained cohort score normalisation) [5, 19, 20]. It should be noted that, in general, a normalised similarity score in speaker verification is expressed as a log-likelihood given by

$$L_{\mathrm{SV}}(\mathbf{O}) = \log p(\mathbf{O} \mid \boldsymbol{\lambda}^{\mathrm{T}}) - \log p(\mathbf{O} \mid \boldsymbol{\lambda}^{\mathrm{I}}), \tag{1}$$

where p indicates probability, $\mathbf{O}$ is the observed test utterance, $\boldsymbol{\lambda}^{\mathrm{T}}$ is the target model (claimed identity) and, $\boldsymbol{\lambda}^{\mathrm{I}}$ is the impostor model which is, in fact, unavailable in practice. As observed, in this formulation $\log p(\mathbf{O} \mid \boldsymbol{\lambda}^{\mathrm{I}})$ provides the normalisation term. In UCN, this normalisation term can be approximated with the average of log-likelihoods for a set of competing speaker models. These competing speaker models are selected from a set of background speaker models based on their closeness to the given test utterance. In practice, it is common to choose the required competing speakers from the set of registered speakers rather than from a separate set. In this case, the normalisation term can be expressed as

$$\Gamma_{\mathrm{UCN}} = (1/K)\sum_{k=1}^{K} \log p(\mathbf{O} \mid \boldsymbol{\lambda}_{\phi(k)}), \tag{2}$$

where $\phi(i) \neq \phi(j)$ if $i \neq j$ and $\boldsymbol{\lambda}_{\phi(1)}, \boldsymbol{\lambda}_{\phi(2)}, \ldots, \boldsymbol{\lambda}_{\phi(K)}$ are the models in the set (other than the target model) which yield the K highest likelihood scores. Figure 1 illustrates the process of unconstrained cohort normalisation in speaker verification.

This way of selecting competing speakers can provide a useful basis for deemphasising the score obtained by each of the twins when targeting the other's reference model. This is due to the fact that, given a sufficiently large background speaker set, the selected competing speakers are the ones that strongly match the combined genetic/non-genetic characteristics in the test utterance. As a result, the uncommon

non-genetic characteristics (e.g. smoking effects) of the twins' voices are implicitly exploited to reduce the score for each speaker targeting the reference model for his or her identical twin.

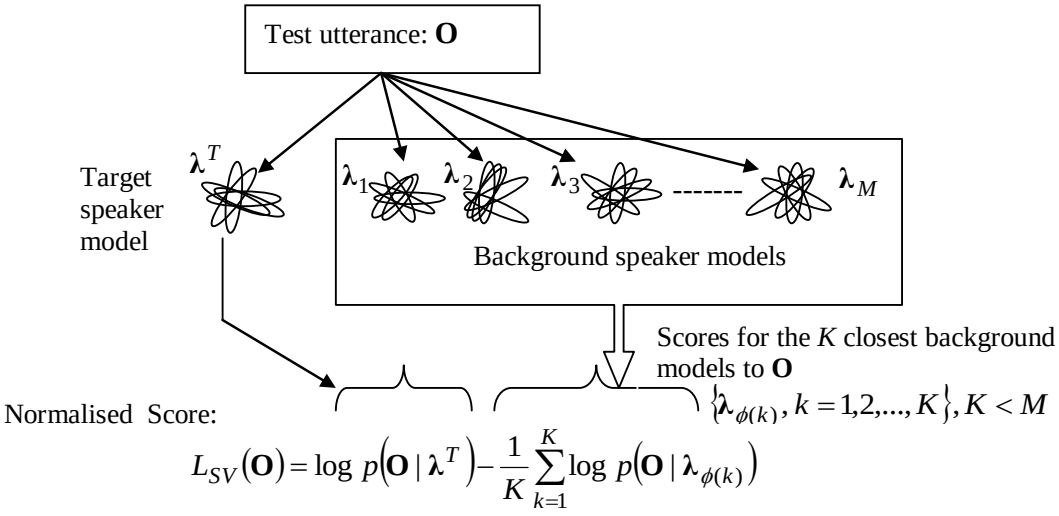

Figure 1: UCN-based score normalisation in speaker verification

MATERIAL AND METHODS

Speech Data

The investigations are based on the use of the only available speech database of identical twins. This database was collected with the support of the Centre for Twin Research and Genetic Epidemiology at St. Thomas' Hospital in London, UK. The data consisted of 49 pairs of identical twins and was dominant in female gender (forty pairs of females and nine pairs of males). From every individual, two token recordings were collected. The first token was a poem, "I wandered lonely as a cloud", by William Wordsworth. This was around 60 seconds in duration. The second token was the date of birth of the individual, spoken as digits. This was around 5 seconds in duration. All the recordings were based on a sampling rate of 44.1 kHz. These were then down-sampled to 16 kHz for the purpose of experiments.

In this investigation, for every individual, the first 30 seconds of the poem data was used to build a reference model. The remaining 30 seconds of the poem data was used for the testing purpose. This is referred to as LONG test data in this paper. The date of birth spoken by each individual is also used for the testing purpose and is referred to as SHORT test data in the remainder of this paper.

Feature Extraction

For the purpose of this study, the t^{th} frame of the input speech data is represented as $c_t \equiv \{[c_t(1), c_t(2),\ldots, c_t(20)],[\Delta c_t(1), \Delta c_t(2),\ldots, \Delta c_t(20)]\}$, where $c(i)$ is the i^{th} static linear predictive coding-derived cepstral (LPCC) parameter and $\Delta c(i)$ is i^{th} delta parameter obtained from the static parameters. The extraction of LPCC parameters is based on first pre-emphasising the input speech data using a first order digital filter and then segmenting it into 20 ms frames at the intervals of 10 ms using a Hamming window.

Speaker Modelling

In this work, the speaker representation is based on the use of adapted Gaussian Mixture Models (GMMs) due to their established effectiveness [1]. The adapted models in this study have 2048 Gaussian components. For the adaptation purpose, a gender independent world model is first obtained by pooling two gender dependant world models. This is created using 100 speakers in the TIMIT speech database. The adapted models are then obtained using a single step Bayesian adaptation procedure [21].

Testing

The verification tests are conducted separately for overall population of speakers, and for the individual pairs of identical twins. These are referred to as the OVERALL and TWIN tests. In the OVERALL configuration, any speaker could claim the identity of any other speaker in the registered population. On the other hand, in the TWIN configuration, each registered speaker can only claim the identity of himself/herself or that of his (her) own identical twin.

With each configuration, the tests are conducted using the SHORT and LONG test tokens. For every test, the results are first obtained using the GMM-UBM scoring procedure. These are used as the baseline results. The scores obtained in this way are then subjected to unconstrained cohort normalisation (UCN), based on a cohort size of 3. The outcomes are referred to as UCN results.

RESULTS AND DISCUSSION

The results obtained for the TWIN and OVERALL configurations are presented in terms of Equal Error Rates (EER %) in tables 1 and 2 respectively. The experimental results for SHORT test tokens are also given as the DET (Detection Error Trade off) plots in Figure 2.

	SHORT	**LONG**
Baseline	10.4	5.2
UCN	1.0	≈0.0

Table 1: Speaker verification performance with and without UCN for the TWIN configuration, in terms of EER (%)

	SHORT	**LONG**
Baseline	2.8	0.4
UCN	0.5	0.0

Table 1: Performance of speaker verification with and without UCN for the OVERALL configuration, in terms of EER (%).

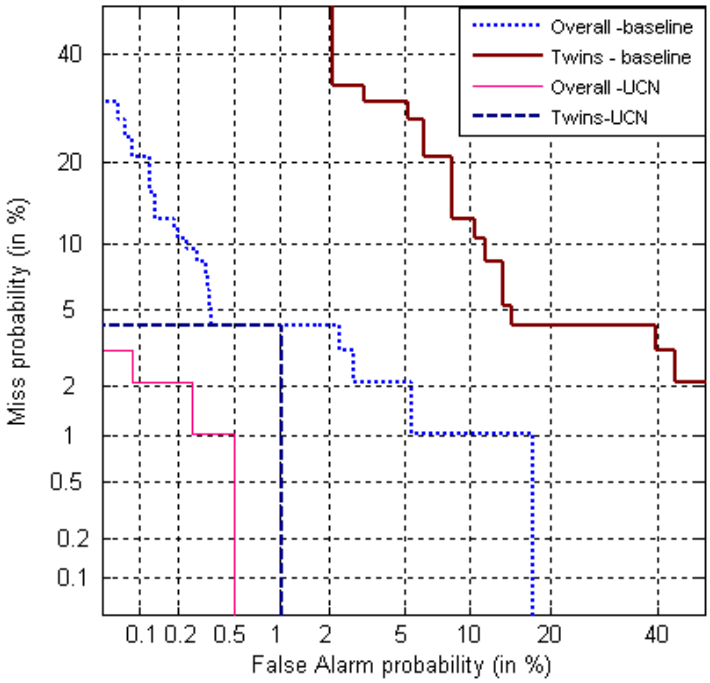

Figure 2: Speaker verification performance with and without UCN for SHORT test tokens

It can be observed from the results in tables 1 and 2 that, as expected, the use of long test utterances leads to smaller error rates. With reference to Table 1 it is noted that, with the TWIN configuration, the EERs are about 10% and 5% in the cases of short and long test tokens respectively. This is a clear indication of the non-genetic (extraneous) factors influencing the characteristics of the voices of each pair of the twins. Without such extraneous effects, the baseline EERs would be expected to be much greater and largely independent of the length of the test utterance used. The use of UCN in this scenario is observed to significantly reduce the EERs. These results are in agreement with the suggested capability of UCN to reduce the impostor scores in relation to those of true speakers. As indicated earlier, in this particular situation, UCN exploits the non-genetic characteristics of the twins' voices to enhance the discrimination

capability of SV. A comparison of the results in Table 2 with those in Table 1 clearly shows that the EERs for the OVERALL configuration are much lower than those for the TWIN configuration. This is caused by the fact that, in the case of the OVERALL configuration, the voice discrimination is based on the genetic as well as non-genetic characteristics of the test utterances. It is observed in Table 2 that again, with this configuration, the use of UCN leads to significant reduction in EERs.

The DET plots for SHORT test tokens in Figure 2 further illustrate the effectiveness of UCN for enhancing discrimination by exploiting the genetic/non-genetic differences between the voices of impostors and target speakers. This capability of UCN can also be observed by examining the score distributions for the true speakers (clients) and impostors in Figure 3. It is observed that the use of UCN leads to considerable reduction in the overlap between the score distribution for clients and those for twin and general impostors. The results for the TWIN configuration clearly demonstrate the significance of exploiting the voice non-genetic characteristics for speaker discrimination.

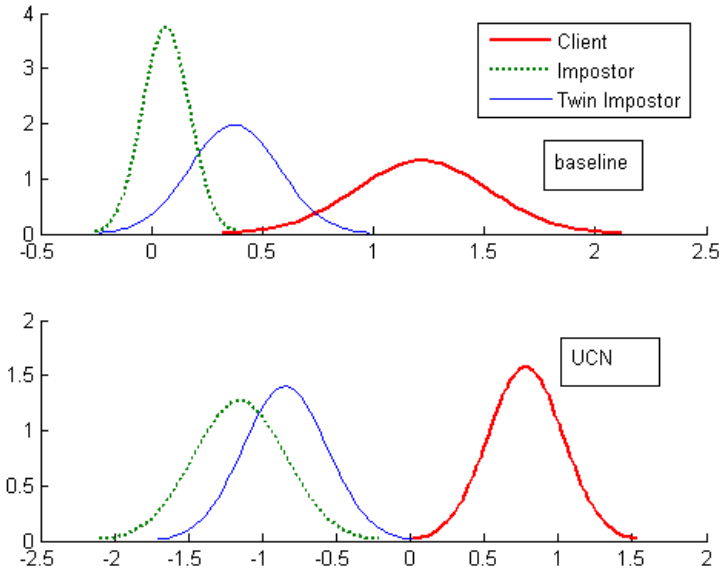

Figure 3: Score distributions for clients, twin impostors and general impostors, obtained using the SHORT test tokens with and without using UCN

CONCLUSIONS

The speaker verification capability for discriminating between identical twins has been investigated. The additional challenge introduced by monozygotic twins in this process is due to their identical physiological developments. However, in cases where the twins experience independence, their voices are subjected to different extraneous influences. This leads to dissimilarities between the non-genetic characteristics of the monozygotic twins. It is shown that, through the use of unconstrained cohort score normalisation (UCN), it is possible to exploit the non-genetic characteristics of the twins' voices for the benefit of increasing the discrimination capability of speaker verification. The experiments with test utterances of about 5 seconds in duration have shown that, with the use of UCN, the EER can be reduced from over 10% to around 1%. Additionally, it has been demonstrated that UCN can also be highly beneficial for exploiting the existing differences in the genetic characteristics of unrelated speakers. This is supported by a set of experiments in which each registered speaker is allowed to target the reference model for his (her) own twin as well as those for other registered speakers. The results have shown that again with the use of UCN the EER can be reduced from around 2.8% to around 0.5% when the test utterances are about only 5 seconds in duration.

ACKNOWLEDGEMENTS

The authors are thankful to the Centre for Twin Research and Genetic Epidemiology at St. Thomas' Hospital, London, for providing help and support in connection with the collection of a speech database for this work.

REFERENCES

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

[2] F. Bimbot, J.-F. Bonastre, C. Fredouille, G. Gravier, I. Magrin-Chagnolleau, S. Meignier, T. Merlin, J. Ortega-Garcia, D. Petrovska and D. A. Reynolds, "A tutorial on text-independent speaker verification", Eurasip Journal on Applied Signal Processing, Special Issue on Biometric Signal Processing, Vol. 2004, No. 4, pp. 430-451, 2004.

[3] R. Auckenthaler, M. Carey, and H. Lloyd-Thomas, "Score normalization for text-independent speaker verification systems", Digital Signal Processing, vol. 10, pp. 42-54, 2000.

[4] K. P. Li and J. E. Porter, "Normalizations and selection of speech segments for speaker recognition scoring", Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'88), vol. 1, pp. 595-598, 1988.

[5] A. Ariyaeeinia and P. Sivakumaran, "Analysis and comparison of score normalisation methods for text-dependent speaker verification", Proceedings of the Eurospeech'97, pp. 1379-1382, 1997.

[6] D. A. Reynolds, "Comparison of background normalisation methods for text-independent speaker verification", Proceedings of the Eurospeech'97, pp. 963-966, 1997.

[7] Dialogues Spotlight Consortium, "Large Scale Evaluation of Automatic Speaker Verification Technology", The Centre for Communication Interface Research, the University of Edinburg, 2000.

[8] A. Cohen and T. Vaich, "On the Identification of Twins by Their Voices", Proceedings of the ESCA Workshop on Automatic Speaker Recognition, Identification and Verification, Martigny, Switzerland, pp. 213-216, 1994.

[9] M. M. Homayounpour and G. Chollet, “Discrimination of voices of twins and siblings for speaker verification”, Proceedings of the 4th European Conference on Speech Communication and Technology (EUROSPEECH ’95), Madrid, Spain, pp. 345-348, 1995.

[10] H. A. Patil and T. K. Basu, “Detection of Bilingual Twins by Teager Energy Based Features”, Proceedings of the International Conference on Signal Processing and Communication (SPCOM), pp 32-36, 2004 .

[11] W. J. Larsen, "Human Embryology," Churchill Livingstone: Philidelphia, 2001.

[12] A. Scheinfeld, “Twins and Supertwins”, Penguin Books Ltd: Middlesex, 1973.

[13] B. M. Carlson, “Human Embryology and Developmental Biology”, 2nd Ed., Mosby: St Louis, 1999.

[14] H. Peeters, S. Van Gestel, R. Vlietinck, C. Derom, and R. Derom, “Validation of a telephone zygosity questionnaire in twins of known zygosity”, Behavior Genetics, vol. 28, 159-163, 1998.

[15] Hunkin J., Personal Communication. St Thomas Hospital, London, 2006.

[16] J. Gonzalez and A. Carpi, “Early Effects of Smoking on the Voice: A Multidimensional Study”, Medical Science Monitor 10(12):CR649-656, 2004.

[17] S. P. Whitesid and E. Rixon, “Speech Characteristics of Monozygotic Twins and a Same-Sex Sibling: An Acoustic Case Study of Coarticulation Patterns in Read Speech”, Phonetica, Vol. 60, No. 4, pp. 273-297, 2003.

[18] F. Nolan and T. Oh, “Identical twins, different voices”, Forensic Linguistics, Vol. 3(1), pp. 39-49, 1996.

[19] J. Fortuna, P. Sivakumaran, A. Ariyaeeinia and A. Malegaonkar, “Relative effectiveness of score normalisation methods in open-set speaker identification”, Proceedings of the Speaker and Language Recognition Workshop (Odyssey), pp. 369-376, 2004.

[20] A. Ariyaeeinia, J. Fortuna, P. Sivakumaran and A. Malegaonkar, "Verification effectiveness in open-set speaker identification," IEE Proceedings Vision, Image and Signal Processing, Vol. 153, Issue 5, pp. 618-624, Oct. 2006.

[21] J. Fortuna, A. Ariyaeeinia, and A. Malegaonkar, "Open-set Speaker Identification Using Adapted Gaussian Mixture Models", Proceedings of the 9th International Conference on Speech Communication and Technology, Lisbon, Portugal, pp. 1997-2000, Sept. 2005.