The International
Journal of
Speech,
Language
and the Law

IJSLL (PRINT) ISSN 1748-8885
IJSLL (ONLINE) ISSN 1748-8893

Article

Automatic speaker recognition of identical twins

Hermann J. Künzel

Abstract

Automatic speaker recognition systems typically rely on parameters derived from resonance features of the vocal tract. This implies that the more similar the geometry of two vocal tracts is, the more similar will be the respective similarity coefficients, or likelihood ratios (LRs). Quite obviously this problem is particularly relevant to related speakers, most extremely for identical (monozygotic) twins. This paper is about an experiment with 9 male and 26 female pairs of identical twins who produced one read and one spontaneous speech sample. An automatic system for forensic speaker recognition (Batvox 3.1) was used to calculate inter-speaker (non-target), (2) intra-twin pair, and (3) intra-speaker (target) LR distributions. Results show that in certain conditions an automatic Bayesian-based system is capable of distinguishing even the vast majority of very similar sounding voices such as those of identical twins. However, the performance of the system used here was superior for male as compared to female voices. Quite obviously the sex-related difference was enhanced by the genetic similarity factor.

KEYWORDS AUTOMATIC SPEAKER RECOGNITION, IDENTICAL TWINS, MONOZYGOTIC TWINS, GENETIC SIMILARITY EFFECT

Affiliation
University of Marburg, Germany
email:kuenzelh@staff.uni-marburg.de

IJSLL VOL 17.2 2010 251–277

doi : 10.1558/ijsll.v17i2.251

equinox
www.equinoxpub.com

equinoxonline

1 Introduction

Identical (monozygotic) twins are known to exhibit the most extreme form of anatomical, physiological and physical similarity among humans. It is even impossible, at least to this day, to distinguish them by DNA analysis, a fact that has already caused problems in producing evidence in a number of forensic cases.[1] In the study of language acquisition by twins it has been observed that the process may evolve so similar and at the same time so particular in both that they may even develop their own speech code (cf. Dodd and McEvoy 1994). Locke and Mather (1989) observed patterns of language and speech acquisition in both monozygotic and dizygotic twins and found that errors in the pronunciation of certain speech sounds (substitutions, omissions, 'distortions'; cf. p. 556) were more frequent in monozygotic twins than in dizygotic twins or unrelated peers. Whiteside and Rixon (2003) investigated characteristics of F2 in consonant-vowel-consonant (CVC) monosyllables of a pair of monozygotic twins and one male sibling and found greater similarities among the twins. Ryalls et al. (2004) introduced the factor of environmental influence on twins' pronunciation. They observed voice onset time (VOT) in two pairs of monozygotic twins. The twins of one pair (age 21) were brought up together while the twins of the second pair (age 70) had been separated at age 25. Afterwards they had been living in the Southeast and Northeast of the United States, respectively. Accidentally, both regional dialects differ in terms of VOT: negative vs. short positive VOT. It was found that the latter pair exhibited much larger VOT differences than the first pair. The authors conclude that source features are more genetically determined than what they call 'filter features', such as VOT and vowel formants, that are more influenced by environmental factors. In a large empirical project with 310 pairs of twins Colledge et al. (2002) tested verbal and nonverbal skills of 4-year olds and measured only a 'moderate' genetic influence on both types of skills. Using data from the same project Kovas et al. (2005) added that both genetic and environmental effects on language skills and deficits were comparable for male and female twins.

As far as the physiological and anatomical structures used for speech production are concerned Lundström (1948) has shown that natural variations between identical twins are much smaller than between non-identical ('fraternal') twins in terms of a host of anatomical parameters related to the jaw and teeth, such as size, breadth, position and inclination of teeth, overbite and others (p. 50f.; cf. also the graphs on pp. 188–191). He concludes: 'On the whole, it appears that genetic factors play an important part, in any case equally important as environmental factors', and '… when extreme malocclusions are concerned […] heredity will be the most important factor' (p. 187). A well known axiom in twin research is that if twins are brought up in the same social

environment (family, school, etc.) they are also exposed to the same conditions for socialisation, including, of course, speech and language acquisition (Galton 1876; Newman et al. 1937). Using the *organic* vs. *learned* dichotomy proposed by Nolan (1983), one may say that in principle the speech of identical twins can be expected to feature the smallest possible amount of inter-speaker variation in both categories, and this is the reason why twins' voices are confused so often by human listeners, including close relatives. However, there is also some evidence that '… speaker specific behaviour is evident phonetically, even between identical twin pairs' (Loakes 2006). The size of the differences inside the same pair of identical twins, even though they are supposed to be much smaller than between non-identical ('fraternal') twins, or just siblings, may vary in terms of the individual (pairs of) speakers and the phonetic parameters that are being analysed (Fuchs et al. 2000; Loakes 2006). In an acoustic phonetic analysis of (only) three pairs of identical twins Nolan and Oh (1996: 48f.) found phonetic differences of different type and size in all three, which lead them to the tentative conclusion '…that identical twins are not necessarily phonetically identical and that they make use of the leeway allowed them by the phonological system'. This hypothesis of idiosyncratic variation in speech production is supported by Johnson and Azara (2000). In a series of listening experiments with identical twins as well as unrelated speakers they found that twins' voices can be distinguished significantly but are more often confused than unrelated speakers (p. 17), and also that 'analysis of the perceptual space for talkers showed that the difference between identical twins was in some cases as large as the difference between unrelated talkers' (p. 2). Using a single pair of monozygotic twins (together with unrelated speakers and professional voice imitators) Rosenberg, in an early comparison of the speaker recognition performance by listeners and 'machine' (in terms of an automatic, acoustic algorithm implemented on a computer), reported that one twin speaker was confused 96 per cent of the time in a same/different listening test. The automatic system, however, was able to distinguish the impostor twin brother 'without error' (Rosenberg 1973: 222).

In principle it can be expected that the distinction of identical twins by their voices will also be a problem for automatic speaker identification (SPID) systems since these are based on parameters of the vocal tract function, most often sets of cepstral coefficients, sometimes amended by a number of derivatives such as deltas or delta-deltas, in order to account for certain features of coarticulation (Gonzalez-Rodriguez et al. 2003; Drygajlo 2007; Ramos-Castro 2007, Przybocki et al. 2007). From the preceding, one can imagine that it is exactly features such as the size and shape of the vocal tract, the complexion of the surface of its inner walls, of the tongue and the lips, that are most similar between identical twins. Grossly speaking, automatic systems exploit the 'sound' of the voice and

disregard the speech content almost completely. It goes without saying that the latter feature is a big asset in the forensic environment since it provides independence of the particular language and allows for voice comparisons across different languages and/or with speech samples in languages that may even be unknown to the forensic speech expert himself. Therefore, automatic SPID systems have become a powerful tool and, with all due caution, a complement to the traditional phonetic-acoustic-linguistic method that is currently being used in courts worldwide (Künzel and Gonzalez-Rodriguez 2003; Rose 2003, 2004). At any rate, the wide-spread use of automatic SPID in certain military applications is a direct consequence of its language independence. The principal aim of the present investigation is to answer the following question: To what degree will the performance of an automatic acoustic forensic SPID system be affected if the target speaker and the reference speaker are identical twins? Put differently: Can an automatic SPID system distinguish between identical twins? At this juncture it should be emphasized that this study was conducted with one particular system. Considering that other systems on the market may differ markedly in performance according to the particular test formats, such as in the NIST evaluations (National Institute of Standards and Technology), one must be cautious to generalise the results.

2 Material and method

2.1 Background

One of the reasons for the scarcity of empirical phonetic studies of the speech behaviour of identical twins is probably the fact that it is difficult to gather data from a large number of twin pairs. In the present case, the investigator was lucky to be asked to participate in a TV production called 'Die Zwillings-Show', in the final round of which a total of 65 pairs of identical twins competed to be 'Germany's most similar twins'. 'Similarity' was to be measured in 'disciplines' such as fingerprint patterns, facial geometry, pain threshold, heart rates before and after physical stress, selecting one's favourite dish, and, among others, voice. In order to quantify voice similarity it was proposed to the producers of the show to employ an automatic SPID system to carry out intra-pair comparisons and use the similarity scores as a criterion. Thus, a ranking for all pairs of twins was obtained and provided to the TV company. In return, the author was granted the right to use for scientific research all the personal data that the participants had submitted to the show, including the voice samples, results from the other 'disciplines' and information such as handedness, body height, weight,

age, sex, diseases, occupation, whether and where they grew up together, etc. With respect to this study it is noteworthy that all participants were also tested positive for monozygoticity before being admitted to the competition. The tests for all 'disciplines' were carried out on a single day and recorded in a large TV studio complex in Cologne, Germany. Parts of these recordings were played back during the live broadcast of the final show.

2.2 Speech material

Due to time constraints induced by the large number of twin pairs that had to undergo the sequence of tests in the twelve 'disciplines' the collection of the speech samples was limited to five minutes per speaker. Later in the day it became obvious that the data collection procedure was running late, so the time allocated to each speaker had to be cut to three minutes. The unfortunate consequence of this was that the full size of the speech sample that had been arranged in the first place was available for 9 male and 26 female pairs only. It consisted of (1) reading the German version of the North Wind & the Sun, which takes between 35–45s; (2) speaking spontaneously for two minutes about one's life as a twin, youth, hobbies, occupation etc. A mobile sound-treated booth had been set up for the recordings inside the large noisy studio hall. The sound signal was recorded on two channels. Channel 1 contained the output of the (same) Sennheiser professional wireless microphone headset worn by all speakers. The other channel recorded the signal picked up by a studio microphone placed on a stand ca. 40cm in front of the speaker. This signal was used as a backup in case the wireless transmission from Channel 1 was degraded. The digital recording, monitoring and mastering were done by sound engineers in the recording centre adjacent to the studio. The final data were made available as WAV files on DVDs at 44 kHz / 32bit.

2.3 Speakers

The TV company had selected twins of all ages, ranging from 7 to 76 years. Even after excluding three children-twin pairs below age 12 for the present investigation the standard deviation of the group mean (34.3 yrs) was still 15.7 yrs. Twins could be identified by a number code given to them by the TV recruiting team. Inside each pair, individuals were identified as 'red' and 'blue'. Pairs with speech defects such as interdental sigmatisms, or click-like noises caused by dentures, or bad reading ability were not discarded since in all these cases both twins were affected.[2] One pair of male twins (M039) who both exhibited diplophonia were also kept in the set. For obvious reasons the kind

and degree of dialectal colouring of the speakers, who came from all parts of Germany, could not be controlled for.

2.4 Pre –processing of speech files

In preparation of the speech data for processing by the automatic SPID system all speakers' audio files were analysed auditorily in order to remove extraneous noise, laughter, simultaneous speaking of subject and interviewer etc. In order to avoid too large a difference in the duration of read and spontaneous samples the latter were limited to ca. 80s including pauses whenever necessary, which resulted in ca. 60s of net speech (since the system eliminates pauses and voiceless speech sounds). Speaking errors in spontaneous or read speech, however, were not removed. Since the automatic SPID system used here was designed to operate in the forensic environment and thus has to cope with telephone-transmitted speech material it requires acoustic data to be sampled at 8 kHz and 16bits.

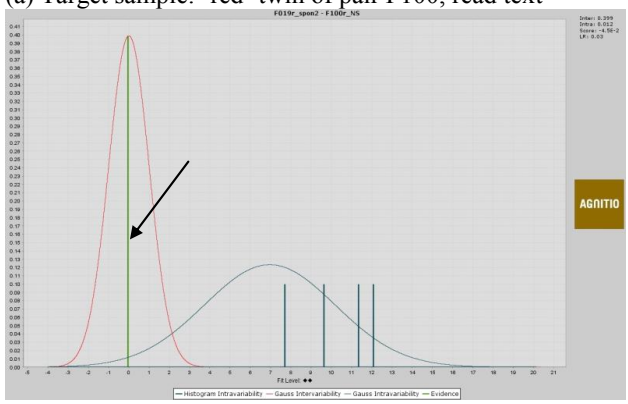
2.5 Automatic speaker recognition

The automatic forensic SPID system used in this investigation is Batvox 3.1[3]. Earlier versions as well as its precursor IdentiVox have been in use since 2000. At this juncture, it is useful to briefly describe its working principle in order to understand the set up of the speech data. Extensive descriptions including the Bayesian foundation, feature extraction, channel normalisation technique and selection of suitable reference populations may be found in Gonzalez-Rodriguez et al. (2003, 2004, 2006). In the LR (Likelihood-Ratio) mode of operation, which corresponds to the typical forensic paradigm of identifying an individual in an open set of individuals, the system matches the voice sample of one or more known (reference) speakers with a sample of an unknown (target) speaker. After passing a signal to noise threshold the reference speaker's sample is automatically split into four sections of equal duration in order to account for intra-speaker variation. A 38-dimensional feature vector calculated every 10 ms is used to represent the resonance behaviour of the vocal cavities of a speaker. It consists of 19 mel-frequency cepstrum coefficients (MFCCs) plus their deltas, which to a certain degree account for coarticulatory features such as transitions between neighbouring speech sounds. The results for the four sections of the reference samples are combined in a statistical model for the reference speaker, which is then compared with the results for the target speaker's sample. The similarity score gained from this procedure is then weighed using

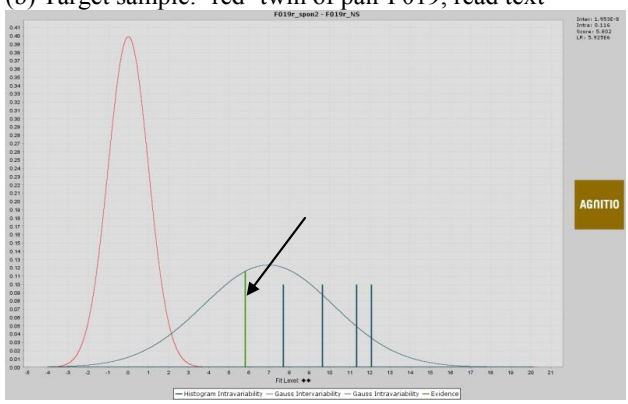
a reference population that can be composed in terms of number of speakers, type of speech material (spontaneous, read, interview etc.) transmission channel characteristics (microphone, landline telephone, GSM, VoIP, analogue radio, TV) and other variables, according to the conditions of the case. The reference population should match the experimental population as closely as possible in terms of parameters such as channel characteristics, language and type of speech material (read, spontaneous etc.). According to the manual of the automatic system used here it should consist of a minimum of 25 subjects. The final measure delivered by the system is a likelihood ratio (LR), with the likelihood for identity of the two voices as numerator ('target and reference voice pertain to the same individual', the so-called prosecution hypothesis) and for non-identity as denominator (defence hypothesis). Technically, the LR is approximated using probabilistic estimations of the reference speaker's variability and the variability of the rest of the speakers 'in the world'. Numerically it is calculated as the ratio of the heights of intra-speaker and inter-speaker distributions at the point of the score that is obtained for the comparison of the target voice with the suspect voice. The LR is 1.0 at the crossing point of both distributions. Consequently, any LR > 1 for an actual no-match would have to be regarded as 'misleading support' of the identity hypothesis, and an LR < 1 for an actual match as 'misleading support' for the contrary. It goes without saying that the higher the LR, the stronger is the support for the respective hypothesis. Since Batvox has been developed for forensic use the normalisation and transformation procedures that eventually turn similarity scores into LRs are based on realistic forensic background data. Since the system is not calibrated for the very small acoustic differences to be expected between identical twins it can be predicted that it will not always yield LRs <1, i.e. it may not always correctly reject the 'blue' sibling of the respective 'red' twin. In any case, the *relative power* of distinction between twins of the same pair can still be measured as the magnitude of the difference between LRs for matches (identical speakers), and intra-twin pair no-matches, i.e. comparisons of 'red' vs. 'blue' twin of the same pair.

With respect to the forensic application another important feature of the system is the option to include so-called case impostors, i.e. speakers who are certainly *not* identical to the speaker under test but exhibit some similarities to the target samples, for instance in terms of channel transmission characteristics. These 'impostors' signify to the system that similarities found between them and the target speaker are not speaker-dependent. Thus the system may recognise certain acoustic resemblances as irrelevant, which enhances its performance. In the forensic environment this mode of operation is often used when in a large

(a) Target sample: 'red' twin of pair F100, read text



(b) Target sample: 'red' twin of pair F019, read text



(c) Target file: 'blue' twin of pair F019, read text

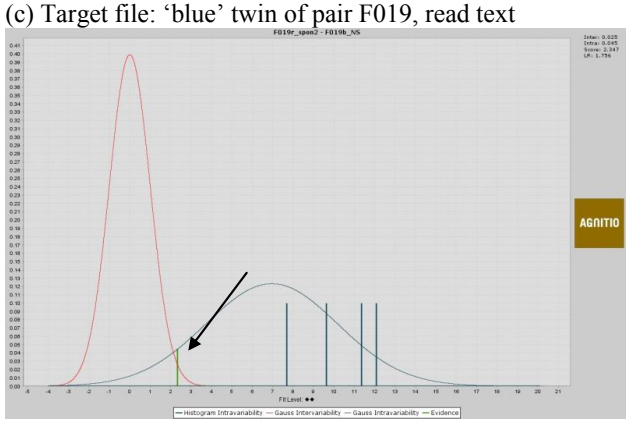


Figure 1: Output of the automatic SPID system

Reference sample = twin speaker F019r, spontaneous speech

case with many speakers some speakers (minimum of three) in some conversations are available who can be excluded as the suspect speaker – for instance the very person(s) that the suspect is talking to. The author's experience with large forensic cases and experiments shows that even when the quality control module of Batvox claims that the reference population chosen for a particular test is acoustically adequate using impostors can still reduce the probability for false acceptance errors, for instance, if all subjects had to read the same text, and/or if the same or a similar microphone and transmission channel were used. In the present set of calculations 5 male and 5 female red-twin speakers were used as case impostors.

A typical example of the performance of the automatic system in the present experiment is illustrated in Figure 1. In Fig. 1a the envelope function on the left side is the distribution of the normalised scores of the members of the reference population (inter-speaker or non-target distribution). At its right slope this function is partially overlapped by the distribution of the reference speaker. The large single bar near the peak of the inter-speaker distribution (indicated by the arrow[4]) represents the score for the comparison of the acoustic model of the target voice. In this case, the bar is very much into the inter-speaker distribution, which is a clear sign of non-identity with the reference speaker. The LR is printed out in the upper right corner of the display (0.012 / 0.399 = 0.03). The figure contains the result for the following comparison: Reference speaker (model) = 'red' twin of pair no. F019 (female), spontaneous speech; target = 'red' twin of pair no. F100, read sample. For this case the system has correctly rejected the prosecution hypothesis. In Figure 1b both reference and target speaker are identical, namely 'red' twin of pair no. F019, with her first 60 seconds of the spontaneous speech sample as reference and her read text as target. Here, the vertical bar (indicated by the arrow) for the target is located clearly inside the intra-speaker distribution at a point where the inter-speaker distribution has asymptotically dropped to almost zero. Accordingly, the LR is quite high (0.116 / 1.953E-8 = 5.9E6) and thus supports the prosecution hypothesis, i.e. both voices are attributed to the same individual (correct identification). Figure 1c contains the result for the match of the spontaneous-speech model of 'red' twin F019, with the read text of her own 'blue' twin sister as a target. Here, the result for the target is located slightly more in the intra-speaker distribution than in the inter-speaker distribution. Since, strictly speaking, any LR larger than 1 means 'identity' this LR of 1.7 would be a 'misleading support' of the identity hypothesis, i.e. a non-distinction of 'red' and 'blue' twin sisters. On the other hand, considering 1) the enormous difference between a LR of 1.7 and the LR of 5.9E6 that was obtained for the non-twin case and 2) the fact that the reference population was far from optimally adjusted to the test (containing

only analogue recordings of unrelated individuals and read speech; see discussion) one would certainly concede that the system was able to distinguish both twins of pair F019.[5]

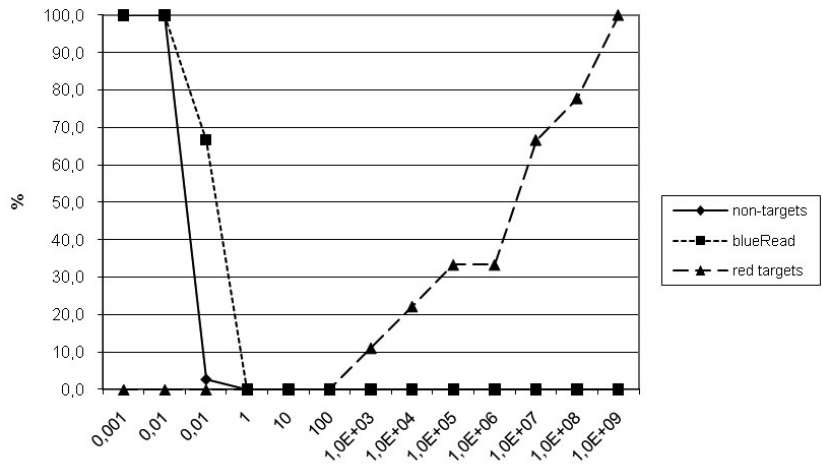
3 Results

3.1 Male twins

For each pair of twins the 'red' brother's first 60s of spontaneous speech were used to build the reference speaker's model. In Test A, the same speaker's read sample and his 'blue' brother's read sample were used as targets. In other words, since both speaking mode and wording of the text were the same for the 'red' and 'blue' target speaker, the only source of acoustic differences between both target voices lies in the speakers. It should be kept in mind, however, that the basis for the 'red' brother's reference model was the spontaneous text, which implies that both speaking mode and wording of reference and targets were different, and furthermore that the material for the reference model was in most cases roughly twice as long as for the target samples since reading of 'The North Wind and the Sun' typically takes only about 35s (whereas the spontaneous samples were ca. 80s long). In Test B the spontaneous speech sample of the respective 'blue' twin was used as target. Although the speaking mode was now the same for model and target the task to detect differences between both twins was by no means easier, since the larger amount of speech material could be expected to contain more of the intra-speaker variability of the target speaker, resulting in more overlap between both twins. Furthermore, the fact that it was not possible to use a sufficiently large reference population consisting of *spontaneous* speech samples a certain amount of mismatch was introduced to the test, whose significance could not be assessed. In both test conditions the same set of 113 male speakers with German as first-language were used as reference population. It consisted of 101 analogue recordings of a fake kidnappers message (ca. 40s, read text; recordings made in 1980 in a quiet office room on a Revox A700 tape recorder) plus the read-text recordings of 12 'red' twin speakers from the present investigation who had produced only one speech sample each and therefore could not be used for the paired comparisons (cf. 2.2).

From now on and for all the experiments the numeric results are expressed as LRs. Strictly speaking, LRs cannot be analyzed using the classical FA/FR (false acceptances/false rejections) approach, but they present a convenient

2a) Distribution of LRs



2b) Tippett plot with data from 2a)

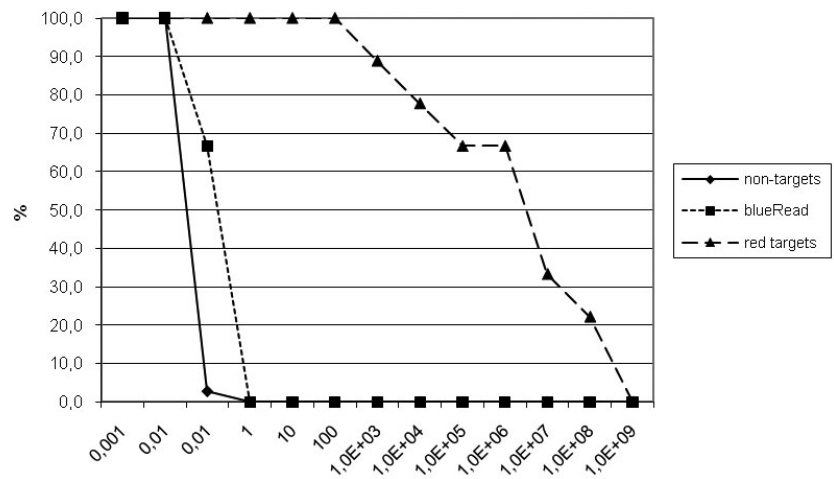


Figure 2: Performance of the automatic SPID system in tests with 9 pairs of identical male twins

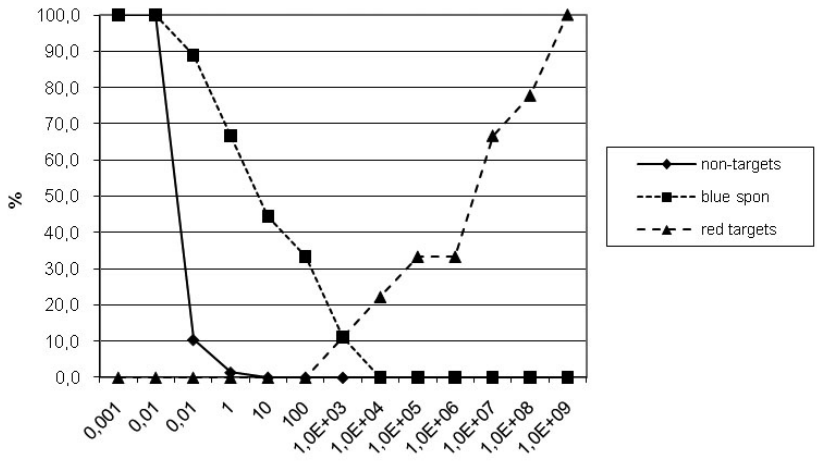
Test A. (1) Solid line: Non-target LRs (i.e. matches of the model of the 'red' twin of each pair with all other 'red' twins' read samples). (2) Dotted line: Intra-twin pair matches, read sample. (3) Dashed line: Intra-speaker LRs (matches of red twin's model (spontaneous) with his own read sample).

way to compare performance across experiments. As such, we use the values as a 'standard score', as well as to extract the EER, thus providing us with a convenient metric to compare performance quickly and easily.

Figure 2 summarizes the results for Test A. In Figure 2a the solid line on the left side indicates the cumulative probability density function (PDF) of LRs for non-target comparisons, i.e. the target voice of each 'red' twin matched with all other 'red' twins' voices (inter-speaker comparisons). The dashed line on the right contains the PDF for target matches (intra-speaker comparisons), i.e. identifications of the ('red') target speakers. Since there is no overlap it would be possible in any case to set a threshold to separate the impostor scores perfectly from the target scores. In this case we can say that the equal-error rate (EER) is zero per cent since there is a working point (threshold) were both errors (FA/FR) are equally zero. The dotted line on the left side is the PDF for intra-pair comparisons, i.e. with the 'red' twin's spontaneous-speech model matched to his 'blue' brother's read sample. The distribution is very close to the non-target PDF and also free of overlap with the intra-speaker function, which means that the system was able to distinguish between both twins in all nine pairs. Figure 2b displays the results as a Tippett plot. It emerges that although the horizontal distance between intra-pair and intra-speaker distributions for the 'red' twin has decreased by a small degree (about one order of magnitude) both are still well separated. The horizontal distance between the non-target and intra-pair functions may be attributed to what may be termed the 'genetic similarity factor'.

The pattern changes significantly for Test B where the target sample of the 'blue' brother also consists of spontaneous speech. In order to facilitate comparisons with the results for Test A, Figure 3 has the same axes and scale as Figure 2, and the intra-speaker results (dotted line) obtained in Test A are copied from Figure 2. In Figure 3a the distributions of inter-speaker and intra-speaker LRs are again free of overlap, i.e. the EER is zero per cent. However, the intra-twin pairs distribution has shifted so far to the right side that it overlaps with the intra-speaker distribution. The crossover point (EER) is 11 per cent (dashed line). Put differently, if the intra-twin pairs distribution was regarded as the non-target distribution the EER for the distinction of identical twins would be 11 per cent. Figure 3b contains the results for Test B as a Tippett plot.

3a) Distribution of LRs



3b) Tippett plot with data from 3a)

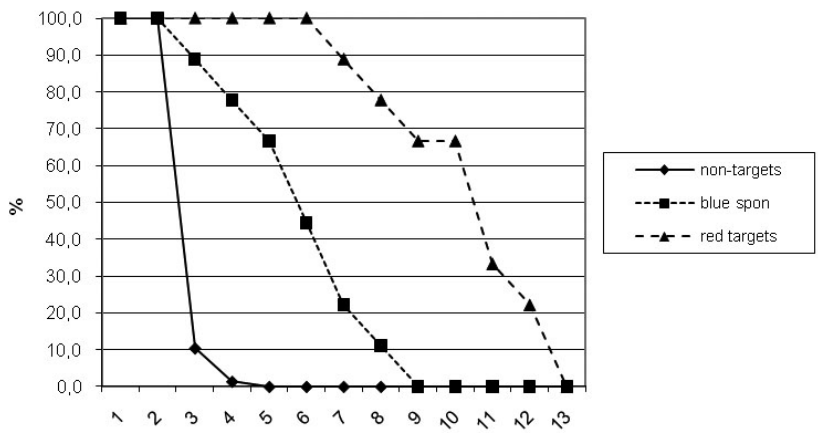


Figure 3: Performance of the automatic SPID system in tests with 9 pairs of identical male twins

Test B. (1) Solid line: Non-target LRs (i.e. matches of the model of the 'red' twin of each pair with the spontaneous samples of all other 'red' twins). (2) Dotted line: Intra-twin pair matches with both twins' spontaneous samples; (3) Dashed line: Intra-speaker LRs (matches of red twin's model (spontaneous) with his own read sample).

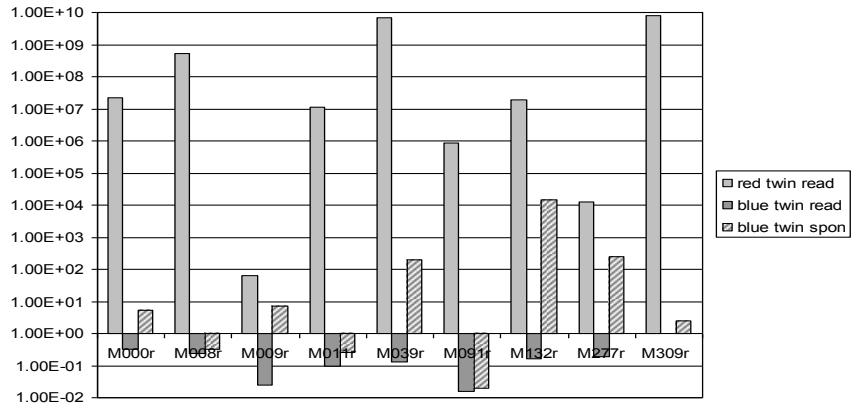


Figure 4: LRs for nine pairs of male identical twins. All models were calculated from the red twin's spontaneous text. Targets (from left to right): 1) Intra-speaker, 2) 'blue' twin's read text, 3) 'blue' twin's spontaneous text

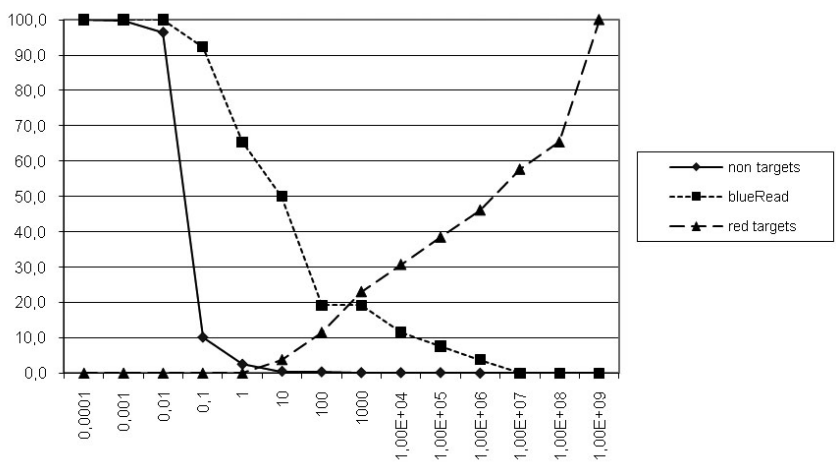
The bar chart in Figure 4 provides a closer look at the data for the individual twin pairs. At first glance the large differences between twin pairs corroborate a finding reported in nearly all studies on speaker identification, be it by man or machine, that some speakers are identified more easily than some others and that a considerable amount of the errors in an experiment may be linked to only a few speakers (cf. Doddington et al. 1998). Each speaker is represented by three columns which represent the following results (from left to right): 1) Identification of the target speakers ('red' twins, read texts); 2) matches of 'red' and 'blue' twins, targets consisting of the read samples (Test A); 3) matches of the 'red' twins with the 'blue' twins' spontaneous speech (Test B). The reference speaker models and the reference population are the same for all three conditions. Correct identifications (left column) appear as LRs larger than unity whereas correct rejections of the twin brother (central and right columns) appear as LRs smaller than unity. The following facts can be stated: 1) In all pairs the 'red' twin is correctly identified (LRs > 1), with 8 of the 9 LRs being larger than 10E4. Furthermore, these LRs are always higher than those for the two no-match conditions. 2) All LRs in the second column (target = 'blue' twin's read text) are smaller than unity, that is, all 'blue' brothers were correctly rejected on the basis of their read speech (the value for M309 is not discernible since it is close to the unity line). Put differently, the twins of each pair were distinguished by the system, albeit by individually differing margins. 3) In the more difficult Test B (spontaneous-speech targets), LRs obtained

for matches with the 'blue' twins are generally larger than those for Test A, and 6 of the 9 surpass the threshold of LR = 1: In a strict sense, these results misleadingly support the identity hypothesis, i.e. they indicate non-distinctions or confusions of 'red' and 'blue' twin. However, considering the still very large distances in relation to the LRs for the left column and the non-optimal reference population one can say that at least in relative terms the twins in all 9 pairs were distinguished by the system.

3.2 Female twins

The speech samples of the 26 female pairs of twins were treated analogously to the samples of the males, the only difference being related to the reference population. It consisted of 76 female speakers, 50 of whom were taken from the analogue voice data bank mentioned earlier for the males (reading a fake kidnapper's message), 15 were second-year linguistics students who had read 'The North Wind and the Sun', and 11 were selected from the set of female red twins from the present investigation who had produced only one speech sample each (i.e. the read sample) and therefore could not be used for the paired comparisons (cf. 2.2).

5a) Distribution of LRs



5b) Tippett plot with data from 5a)

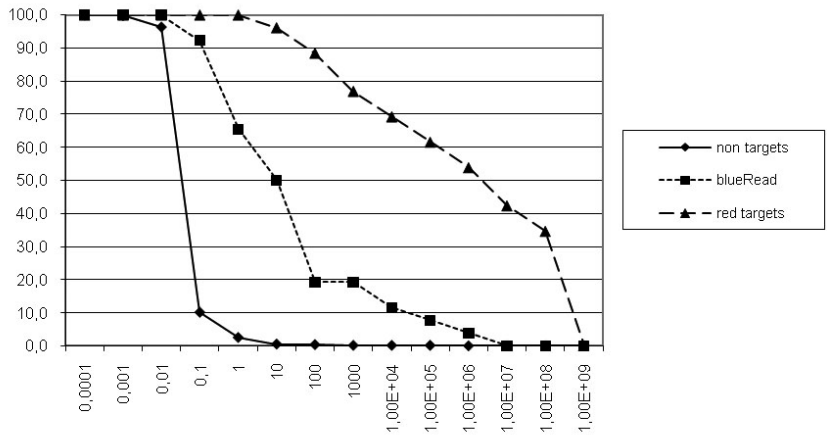


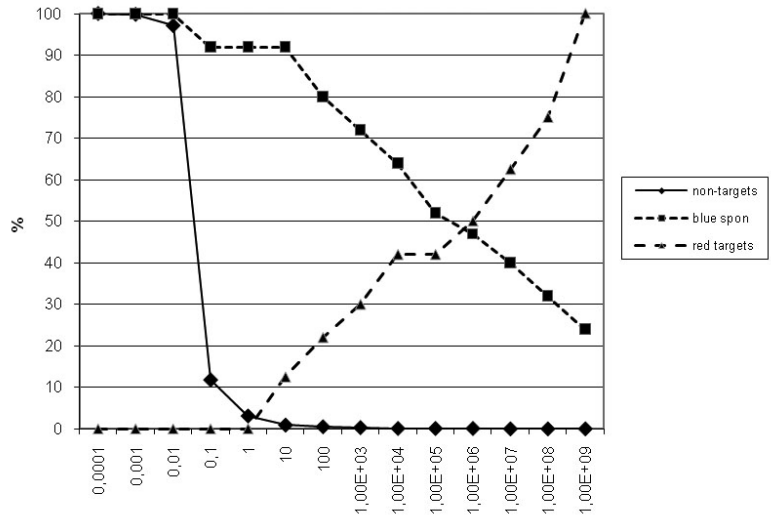
Figure 5: Performance of the automatic SPID system in tests with 26 pairs of female identical twins

Test A. (1) Solid line: Non-target LRs (i.e. matches of the model of the 'red' twin of each pair with all other 'red' twins' read samples). (2) Dotted line: Intra-twin pair matches, read sample. (3) Dashed line: Intra-speaker LRs (matches of red twin's model (spontaneous) with her own read sample).

Figure 5a contains the PDFs for inter-speaker, intra-twin pair and intra-speaker comparisons for Test A. The PDFs for inter and intra-speaker comparisons show a slight overlap at a LR range between 1 and 10, which corresponds to an EER of 0.5 per cent. This result indicates that the 26 female red twins were nearly always correctly identified when their speaker model was built of the spontaneous sample and the target file consisted of the read sample. When the 'blue' twin sisters' read samples were tested against the spontaneous-speech model the EER was 19.2 per cent. The PDF for the LRs of the 'blue' twins is shifted to the right and intersects with the intra-speaker PDF at an LR range between 100 and 1000. In sum, the performance of the system for female voices is inferior compared to male voices and the fact that the crossover point has shifted so much to the right indicates that under the conditions of this experiment the system as a whole was less well adapted ('calibrated') to cope with female twin voices. Figure 5b contains the data as a Tippett plot. The genetic similarity effect in terms of the horizontal distance between inter-speaker and intra-twin pair distributions is larger than for the male subjects (cf. Fig. 2a) and the distance to the intra-speaker distribution is smaller, although still between two and six orders of magnitude.

As could be expected, the performance of the system was worse in Test B where intra-speaker variability was increased by using the spontaneous sample of the 'blue' twin sisters as targets. Figure 6a shows that the crossover point for

6a) Distribution of LRs



6b) Tippett plot with data from 6a)

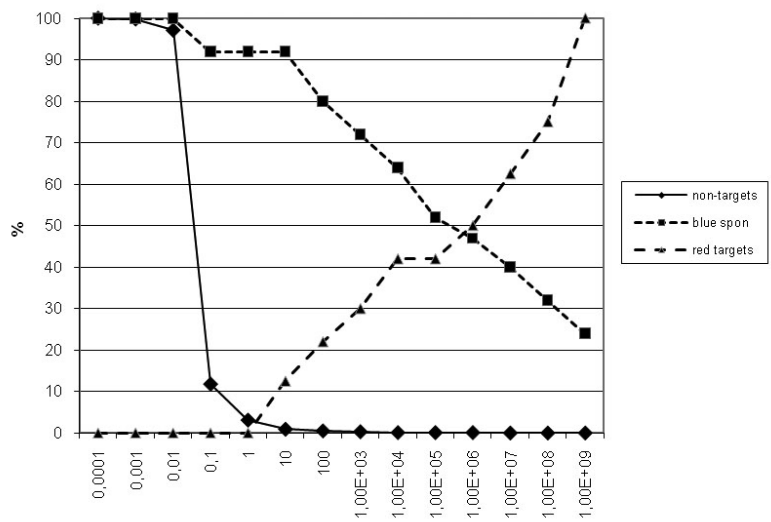


Figure 6: Performance of the automatic SPID system in tests with 26 pairs of female identical twins

Test B. (1) Solid line: Non-target LRs (i.e. matches of the spontaneous text of the 'red' twin of each pair with the spontaneous samples of all other 'red' twins). (2) Dotted line: Intra-twin pair matches with both twins' spontaneous sample. (3) Dashed line: Intra-speaker LRs (matches of red twin's model (spontaneous) with her own read sample).

the inter speaker / intra speaker distributions has increased to 4.4 % and the PDF for intra-twin pair comparisons has shifted far to the right and intersects with the intra-speaker PDF ('red' targets, read samples; same as for Test A) at a LR range between 1.0E5 and 1.0E6. Here, the crossover point is 48 per cent. In the Tippett plot (Figure 6b) it is clearly visible that the intra-pair function is very close to, and even intersects, the intra-speaker distribution, indicating that non-distinctions between female twins are almost as likely as distinctions. Figure 7 contains all the results for the individual pairs. The data are arranged in the same way as those for the male twins (cf. Figure 4). The following observations can be made. 1) In all 26 cases the red twin was correctly identified on the basis of her read speech sample, with individual LRs varying between 15 and 1.0E10. 2) In Test A, where both target voices consisted of the same (read) text, LRs smaller than unity were obtained for 8 of the 'blue' sisters, i.e. they were clearly correctly distinguished from their siblings. In the majority of cases, however, LRs larger than unity were obtained for the 'blue' sisters. Strictly speaking, these are incidences of false acceptance, i.e. non-distinctions of both female twins. However, all LRs for the comparisons with the 'blue' twins are still much lower than the LRs for the identification of their 'red' sisters. 3) In Test B, however, only 2 'blue' twin sisters scored LRs lower than unity and thus were distinguished from their red sisters in absolute terms. Another 13 'blue' twins obtained positive LRs that were, however, lower than the corresponding values for their 'red' sisters. In these cases both twins were distinguished in relative rather than absolute terms. A total of 11 'blue' sisters obtained equal or even higher LRs than their own 'red' sisters. In these cases, both twins were clearly not distinguished by the automatic system.

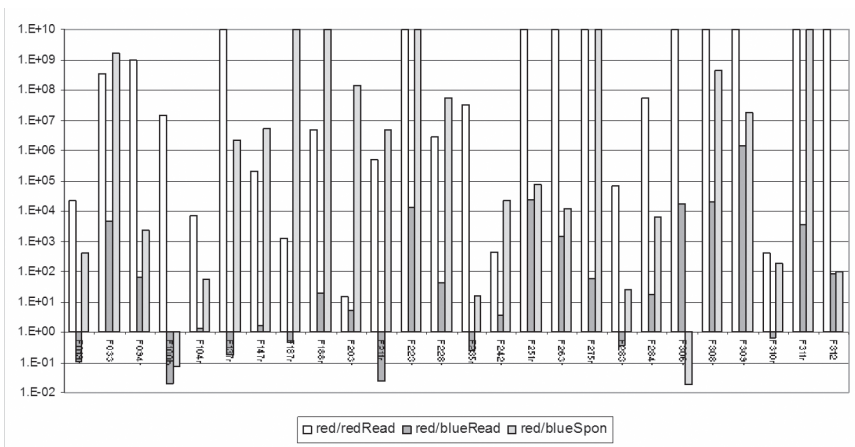


Figure 7: LRs for 26 pairs of female identical twins. All models were calculated from the 'red' twin's spontaneous text. Targets (from left to right): (1) intra-speaker (white), (2) blue twin's read text (dark grey), blue twin's spontaneous text (grey).

4 Discussion

4.1 The factor of genetic similarity

Perhaps the only entirely uncontroversial finding in numerous studies involving machine speaker identification, acoustic-phonetic methods, and also in the various types of perceptual tests, is that some voices usually involve more errors than others in terms of false acceptances or false rejections (cf. the terminology and discussion in Doddington et al. 1998). Using identical twins to test the performance of an automatic SPID system may be considered as a most challenging task since the genetic similarity implies the strongest possible reduction of between-speaker variation. Put differently, what may be called the 'genetic similarity factor' simply means that the *a priori* chances for a target voice to be *very* similar to the reference voice is much larger than within a set of unrelated individuals.

Regarding the performance of the SPID system equal error rates obtained for *unrelated* speakers are low: zero per cent for the males and 0.5 per cent for the females. To be fair, it must be acknowledged that unlike the usual forensic setting the speech signals used for this investigation were direct recordings of good quality and exhibited the same transmission channel characteristics. Furthermore, the recommended minimum requirement by the system of 60s of speech for the calculation of a reliable speaker model was fulfilled in all cases. On the other hand, there was some mismatch in terms of the reference populations that had to be used. Populations for tests with male and female voices consisted mainly of old analogue recordings (made in 1980) plus a few digital recordings, and all contained only read speech. Perhaps more importantly, all these recordings were shorter than 1 minute (35–50s). Since a reference population should match the features of the target voices as closely as possible the pre-requisites were best for Test A, where targets consisted of short stretches of read speech. Since no directly-recorded reference population with *spontaneous* speech was available, the same reference population had to be used for Test B, which created another mismatch with the spontaneous targets in terms of speaking mode and amount of speech material.

4.2 Speaker sex, type and duration of voice samples

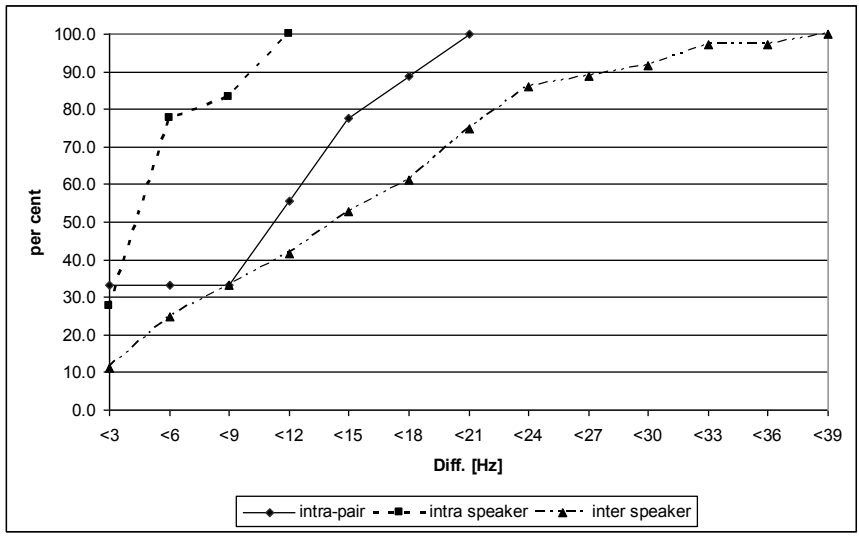
Returning to the central question of this investigation, i.e. whether an automatic SPID system can distinguish identical twins by their voices, it has to be acknowledged that the genetic similarity of twins does indeed affect the ability of an automatic SPID system to recognise their voices in the sense that a correct rejection, or distinction, of the *other* twin is made more difficult. When the target samples for both twins consisted of read speech and identical wording

all 9 male twins were correctly identified *and* distinguished from their twin brothers. When the target samples pertained to different modes of speech, which under the terms of this experiment also means differences in terms of amount of speech material, then the rate of false acceptance of the other twin was 11 per cent. In relative terms, however, both twins in all 9 pairs were still correctly distinguished. Results were not as good for female twins. Under the favourable condition of same-text target samples for both twins the equal-error rate was 19.2 per cent. It increased to 48 per cent when both speaking mode of the target samples and amount of speech material were different, which means that in the latter condition twin sisters could be distinguished within only half of the pairs.

At this juncture, two important questions have to be discussed. First: What is the reason for the large sex-related differences? Second: Why does a higher amount, and different mode, of speech lead to a degradation of the performance of the automatic system? In fact, it might be the case that both results are related. Trying to interpret the sex-related discrepancy of the results reminds one of findings that go back to the pioneers of sound spectrography in the 1950s who found that female voices are generally more difficult to analyse in terms of spectrum-related parameters such as formant centre frequency and bandwidth: 'One of the greatest difficulties in estimating formant frequencies was encountered in those cases where the fundamental frequency was high' (Peterson and Barney 1952: 181; see also House 1959). In the current investigation, the mean values for fundamental frequency are 125 Hz for the males and 220 Hz for the females. As a consequence of the higher fundamental frequency of female voices the spacing of the harmonics is less dense than for male voices, which in turn yields less speech sound- and speaker related information in the spectrum. The same would apply to the extraction of coefficients from the cepstrum. Apart from the higher Fo, female speakers normally also have smaller vocal cavities than males, which leads to higher formant[6] frequencies altogether (Ladefoged and Broadbent 1957, Coleman 1971). As far as the automatic SPID system used here is concerned the sex-related effect is normally quite small. When the results obtained in the latest NIST evaluation (NIST SRE 08) are separated for male and female speakers the EERs are 5.3 and 7.0 per cent for telephone speech. These values are comparable to those for the unrelated speakers in the present investigation (0 and 4.4 per cent, direct recordings).

The fact that the genetic similarity reduces inter-speaker differences makes speaker distinction more difficult. On the other hand, if the target speech samples pertain to the same speaking mode, or even the same wording, such as

a) male speakers



b) female speakers

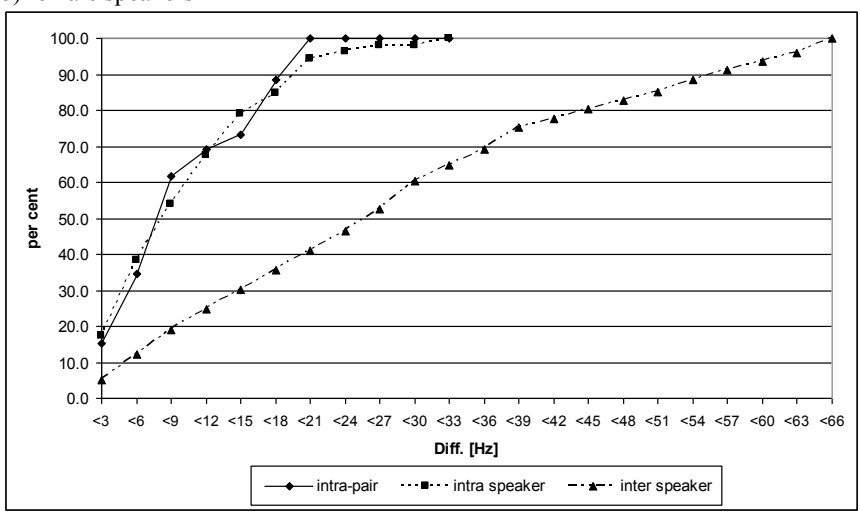


Figure 8: Cumulative distribution of intra-speaker, intra-twin pair and inter-speaker differences of mean Fo values for 9 male and 26 female pairs of identical twins

in Test A of this experiment, then the intra-speaker variability is also reduced to a minimum, which makes the recognition task easier. In a way, one factor may, to a certain degree, offset the other. If, in turn, both target speech samples exhibit differences in terms of overall duration and speaking mode then the aggravating effect of genetic similarity cannot be mitigated. Rather, another adverse effect arises. Since the task of recognition, and in particular correct rejection, of similar-sounding voices can generally be considered to be more difficult for female voices it is obvious to assume that here the combination of these two adverse effects may degrade the performance of an automatic system even more.

4.3 Are women's voices more problematic?

Nevertheless, it remains doubtful whether these two effects alone can explain the large *size* of the loss of performance of the automatic SPID system for the distinction of female twins. Considering the acoustic peculiarities of high-pitched voices discussed earlier the question arises whether it is simply possible that the voices of female twins are *eo ipso* more similar, or less distinct, than those of male twins. This hypothesis will be verified through a follow-up test with the same speech material presented to human listeners in a same / different discrimination test. The receiver operating characteristics (ROC) curves will then reveal any sex-related effect. As a matter of fact, in preparation of the perception test the fundamental frequency values for male and female twins have already been calculated. Figures 8a,b contain the cumulative distributions of intra-speaker, intra-twin pair and inter-speaker differences of Fo for male and female twins. It emerges that the data for both sexes are quite different and support the above hypothesis of greater similarity of female twins' voices: For male twins there is a clear separation of intra-speaker and intra-twin pair Fo-variation: As could be expected, intra-speaker variation (between the read and spontaneous texts) is smallest. 50 per cent of all differences are 4.5 Hz or less. The respective value for intra-twin pair differences is 11 Hz, and 14 Hz for inter-speaker differences. The picture changes drastically for the female subjects (Figure 8b). Intra-speaker variations and intra-twin pairs distribution overlap almost completely, the 50 per cent values for both being 8.8 Hz, and both distributions are well separated from the inter-speaker distribution that contains a 50 per cent value of 26 Hz. Thus one can say that at least with regard to average Fo, which is known to be an important parameter in speaker recognition by listeners, it would be almost impossible to distinguish both twins of the female pairs used in this study.[7]

4.4 Future work

A more general issue for the perception test is to investigate how the performance of the automatic SPID system compares with the ability of human listeners to distinguish between the voices of identical twins. Informal pre-tests suggest that auditory distinction will involve higher error rates, even when the same test sentences are used for both twins of one pair and when they are presented in a same-different format, which is the most sensible test to detect differences. Another desiderate is to investigate the effect of telephone transmission and other forms of acoustic degradations on the performance of the automatic system. This situation may be compared to the task of visually identifying identical twins on a photo that has been blurred on purpose. Furthermore, the comparisons undertaken in this investigation will be repeated on the basis of an improved reference population that also contains spontaneous speech samples. This will help mitigate the mismatch problem that undoubtedly contributed to the right-shift of the intra-twin pair LR distributions.

5 Concluding remarks

The results of the present investigation corroborate previous findings that as far as the speech behaviour is concerned even monozygotic twins cannot generally be considered to be exact copies in terms of voice and speech (Nolan and Oh 1996, Johnson and Azara 2000, Loakes 2006, Whiteside and Rixon 2003). The early medical-statistical dissertation by Lundström (1948) provides some of the anatomical and physiological bases for this finding. From a biological point of view it has been emphasized that there are three different types of monozygotic twins that can be distinguished already in the embryonic and fetal stages (see Loakes 2006: 37–39 for details). Perhaps the most promising discovery with regard to the individuality issue has been made recently in genetics, namely '… the influence of chemical modifications of the DNA on phenomena observable in cell contents as well as behaviour of the cell within its tissue' (Schmitter 2004: 129): Throughout morphogenesis the DNA of both twins is subjected to chance effects of a chemical process called methylation of the cytosine, which acts like an on-off switch on certain parts of the DNA. This effect would also explain why sometimes only one monozygotic twin develops a certain disease.

Speaker identification of identical twins will never become a standard task in the forensic environment, the most trivial reason being the low incidence of one in 250 births in Western Europe (Dudenhausen 2003: 301). However,

the extreme form of genetic similarity that also extends to the structures used for speech production may help develop a benchmark for the performance of speaker recognition methods, especially automatic SPID systems. A system that identifies an identical twin without falsely accepting the other twin is probably fit for use in the forensic environment.

Notes

1 In a recent child maintenance case in Germany, plaintiff P asserted that his twin brother T was the father of the child whom P's wife had given birth to. Both the wife and T admitted to have had intercourse during the period of conception of the child. The court found it impossible to determine by any scientific method which twin brother was the biological father of the child. By default, P was declared the legal father since the child had been born in his marriage. With a slightly ironic touch one could say that from a Darwinist point of view the whole question is irrelevant because the biological information passed on to the next generation by either twin father is identical anyway.

2 People who do not personally know monozygotic twins can hardly imagine how similar they usually are, not only in terms of their physique but each and every aspect of their behaviour, in particular speech. In the free speech recordings made for this investigation many or even most of the of twins – although all of them were recorded single in the sound-treated cabin – would use exactly the same words and phrases and even whole sentences, and in the same order. They would make (the same phonetic type of) pauses inside and between phrases at the same syntactical positions, and produce the same slips of the tongue in the same words. The twins usually had the same occupation and hobby, some had also married other twins, and many had got married, and even divorced, in the same year. When speakers were asked to deliver the two minutes of free speech most of them would use the plural straight away, whereas others would ask the interviewer 'Do you want me to say 'I' or 'we' when I talk?'

3 Manufactured by Agnitio S.A., Madrid, Spain. For detailed information readers are referred to the following document: www.agnitio.es/ingles/files/BATVOXpresentation.PDF.

4 The original graphical display of the results is in colours, with the vertical bar for the target sample appearing green and the bars for the four sections of the reference sample appearing blue.

5 Evett (1998: 201f.) recommends that in all forensic disciplines except DNA 'we must use linguistic qualifiers' to indicate to the court the level of support that a LR gives to the stated propositions. He proposes the following grades (Table 1, p 201): LR 1 to 10: 'limited support', LR 10 to 100:

'moderate support', LR 100–1000: 'strong support', LR > 1000: 'very strong support'. On the basis of personal experience with automatic speaker ID the current author would not consent to use such a scale generally, that is in all cases, until a more quantifiable and objective way to assess the fitness of a reference population to a given case has been found.

6 It is actually more appropriate to speak more generally of *resonance* frequencies since it has been shown that speaker sex can be identified even on the basis of voiceless obstruents such as /s, S / (Ingeman 1968; Schwartz 1968).

7 One reviewer of this paper has pointed out that this statement is based only on the *distributions* of Fo values and '… what we don't know is within each twin pair how similar the intra-speaker and intra-twin differences are.' A follow-up analysis of this question shows the following pattern. Comparing the 26 'blue' + 26 'red' intra-speaker Fo values with the respective 26 intra-pair values it emerges that the number of cases where intra-speaker variation is *smaller* than intra-pair variation is 25, and the same number was found for intra-speaker variation being *larger* than intra-pair differences. The remaining 2 cases (one 'red' and one 'blue' case) exhibit the same values for intra-speaker and intra-twin pair Fo. Put differently, the almost complete overlap of Fo variation for female identical twins can be found also within each twin pair. With regard to Figure 8 and the finding that female twins' voices are more similar than male twins' voices it is not surprising that the picture is completely different for the male twins. Comparing the 9 'red' and 9 'blue' intra-speaker Fo values with the respective 9 intra-twin pair values, 17 of the 18 cases show smaller intra-speaker differences.

References

Agnitio Voice Biometrics (2009) Batvox 3.0 Basic User Manual. Available at *www.agnitio.es/ingles/contacto*

Coleman, R. O. (1971) Male and female voice quality and its relationship to vowel formant frequencies. *Journal of Speech and Hearing Research* 14(3): 565–577.

Colledge, E. and Bishop, D. (2002) The structure of language abilities at 4 years: a twin study. *Developmental Psychology* 38(5): 749–757.

Dodd, B. and McEvoy, S. (1994) Twin language or phonological disorder? *Journal of Child Language* 21: 273–289.

Doddington, G., Liggett, W., Martin, A., Przybocki, M. and Reynolds, D. A. (1998) Sheep, goats, lambs and wolves: a statistical analysis of speaker performance in the NIST 1998 speaker recognition evaluation. *Proceedings of ICSLP 37*: 122.

Drygaijlo, A. (2007) Forensic automatic speaker recognition. *IEEE Signal Processing Magazine* 24: 132–135.

Dudenhausen, J. W. (2003) Die Mehrlingsschwangerschaft. In H. Bender, K. Diedrich and W. Künzel (eds) *Handbuch der Frauenheilkunde und Geburtshilfe* vol. 7: 301–309. München: Urban & Fischer.

Evett, I. W. (1998) Towards a uniform framework for reporting opinions in forensic science framework. *Science and Justice* 38(3): 198–202.

Friedrich, W. (1986) Zwillingsforschung. Ein Überblick über ihre Ansprüche und Probleme. In W. Friedrich und O. Kabat vel Jov (eds) *Zwillingsforschung International* 13–18. Berlin: Deutscher Verlag der Wissenschaften.

Fuchs, M., Oeken, F., Hotopp, T., Täschner, R., Hentschel, B. and Behrendt, W. (2000) Die Ähnlichkeit monozygoter Zwillinge hinsichtlich Stimmenleistungen und akustischer Merkmale und ihre mögliche klinische Bedeutung. *HNO* 48(6): 362–469.

Galton, F. (1876) The history of twins as a criterion of the relative powers of nature and nurture. *Journal of the Anthropological Institute of Great Britain and Ireland* V: 391–406.

Gonzalez-Rodriguez, J., Drygajlo, A., Ramos-Castro, D., García-Gomar, M. and Ortega-García, J. (2006) Robust estimation, interpretation and assessment of likelihood ratios in forensic speaker recognition. *Computer Speech and Language* 20: 331–355.

Gonzalez-Rodriguez, J., Fierrez-Aguilar, J. and Ortega-Garcia, J. (2003) Forensic identification reporting using automatic speaker recognition systems. *Proceedings IEEE – ICASSP* vol. 2: 93–96.

Gonzalez-Rodriguez, J., Ramos-Castro, D., García-Gomar, M. and Ortega-García, J. (2004) On robust estimation of likelihood ratios: the ATVS-UAM system at 2003 NFI/TNO forensic evaluation. *Proceedings of Odyssey 04 Speaker and Language Recognition Workshop, Toledo, Spain*: 83–90.

House, A. S. (1959) A note on optimal vocal frequency. *Journal of Speech and Hearing Research* 2(1): 56–60.

Ingeman, F. (1968) Identification of the speaker's sex from voiceless fricatives. *Journal of the Acoustical Society of America* 44(4): 1142–1144.

Johnson K. and Azara M. (2000) The perception of personal identity in speech: evidence from the perception of twins' speech. Unpublished manuscript, available at: citeseer.ist.psu.edu/ 467059.html

Kovas,Y., Oliver, B., Dale, P., Bishop, D. and Plomin, R. (2005) Genetic influences in different aspects of language development: the etiology of language skills in 4.5-year-old twins. *Child Development* 76(3): 632–651.

Künzel, H. J. and Gonzalez-Rodriguez, J. (2003) Combining automatic and phonetic-acoustic speaker recognition techniques for forensic applications. *Proceedings 15th Intern. Congress of Phonetic Sciences.* Barcelona, Spain. 1619–1622.

Ladefoged, P. and Broadbent, D. (1957) Information conveyed by vowels. *Journal of the Acoustical Society of America* 29: 98–104.

Loakes, D. (2006) A forensic phonetic investigation into the speech patterns of identical and non-identical twins. Unpublished PhD Dissertation, University of Melbourne.

Locke, J. and Mather, P. (1989) Genetic factors in the ontogeny of spoken language: evidence from monozygotic and dizygotic twins. *Journal of Child Language* 16: 553–559.

Lundström, A. (1948) *Tooth size and occlusion in twins.* Basel: Karger.

Newman, H. H., Freeman, F. N. and Holzinger, K. J. (1937) *Twins. A Study of Heredity and Environment*. Chicago: University of Chicago Press:.

NIST (2008) *The NIST Year 2008 Speaker Recognition Evaluation Plan*. nist.gov/speech/tests/sre/2008/

Nolan, F. (1983) *The Phonetic Bases of Speaker Recognition*. Cambridge: Cambridge University Press.

Nolan, F. and Oh, T. (1996) Identical twins, different voices. *Forensic Linguistics* 3(1): 39–49.

Peterson, G. E. and Barney, H. L. (1952) Control methods used in a study of the vowels. *Journal of the Acoustical Society of America* 24(2): 175–184.

Przybocki M. A., Martin A. F. and Le, A. N. (2007) NIST speaker recognition evaluations utilizing the Mixer corpora 2004, 2005, 2006. *IEEE Transactions on Audio, Speech and Language Processing* 15(7): 1951–1959.

Ramos-Castro, D. (2007) Forensic evaluation of the evidence using automatic speaker recognition systems. Unpublished PhD Dissertation, Universidad Autónoma de Madrid.

Rose, P. (2003) The technical comparison of forensic voice samples. In I. Freckelton and H. Selby (eds) *Expert Evidence* Issue 99. Sydney: Lawbook.

Rose, P. (2004) Technical speaker identification from a Bayesian linguist's perspective. *Odyssey-04, ISCA Speaker and Language Recognition Workshop.* Toledo, Spain.

Rosenberg, A. E. (1973) Listener performance in speaker verification tasks. *IEEE Transactions on Audio and Electroacoustics* 21(3): 221–225.

Ryalls, J., Shaw, H. and Simon, M. (2004) Voice onset time production in older and younger female monozygotic twins. *Folia Phoniatrica* 56: 165–169.

Schmitter, H. (2004) The future of DNA analysis from the view of biometrics. In Bundesamt für Sicherheit in der Informationstechnik (ed.) *2nd BSI Symposium on Biometrics* 119–122. Bonn: SecuMedia.

Schwartz, M. F. (1968) Identification of speaker sex from isolated, voiceless fricatives. *Journal of the Acoustical Society of America* 43(5): 1178–1179.

Whiteside, S. P. and Rixon, E. (2003) Speech characteristics of monozygotic twins and same-sex siblings: an acoustic case study of coarticulation patterns in read speech. *Phonetica* 60: 273–297.