The International Journal of Speech, Language and the Law

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

Article

Automatic speaker recognition with cross-language speech material

Hermann J. Künzel

Abstract

Automatic systems for forensic speaker recognition (FASR) claim to be largely independent of language based on the fact that feature vectors are composed of acoustic parameters that are derived from the resonance characteristics of vocal tract cavities. Yet a certain 'language gap' may remain which may deteriorate the performance of a system unless properly compensated. This forensic aspect of what may be called cross-language speaker recognition has not yet received due attention. Based on the most common forensic cross-language setting, the aim of this study was to assess the effect of language mismatch on the performance of a standard FASR system and compare its magnitude with the effect of other sources of mismatch on the same voice data. Using the automatic system Batvox 3 in an experiment with 75 bilingual speakers of seven languages and four kinds of transmission channels, it can be shown that, if speaker model and reference population are matched in terms of language, the remaining mismatch between speaker model and test sample can be neglected, since equal error rates (EERs) for same-language or cross-language comparisons are approximately the same, ranging from zero to 5.6%. Transmission of the speech data via landline telephone, GSM and, for part of the corpus, VoIP (using Skype) caused EERs to rise by less than 1% on average.

KEYWORDS FORENSIC SPEAKER RECOGNITION, AUTOMATIC SPEAKER RECOGNITION, CROSS-LANGUAGE SPEECH MATERIAL, TRANSMISSION CHANNEL CHARACTERISTICS

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

IJSLL VOL 20.1 2013 21–44 doi : 10.1558/ijsll.v20i1.21

Introduction

In the forensic practice of this author the majority of speaker-recognition cases involve speech material from more than one language. Here is a typical example: West-African defendant D was indicted for trading with illicit drugs. Incriminating telephone conversations in Igbo intercepted by the police revealed details of the deals. D conceded that the GSM phone to which the conversations had been tracked may have been his but that it had been stolen from him a few days before the calls were recorded. He also insisted that the only languages he was able to speak were (Nigerian) English and some German. Since he refused to provide a speech sample in either language, which is possible in Germany due to the *nemo tenetur* principle, the court ruled that the expert use as reference material several calls in German 'made undoubtedly by the defendant' to the city's social welfare department, using his real name, in order to enquire about the conditions for receiving social benefits.

Automatic speaker-recognition systems for military and intelligence applications have had to cope with the cross-language problem for several years, especially when large-scale real-time monitoring of communication channels is the task. The standard scenario here is that, for instance, one million telephone or wireless conversations have to be checked per day as to whether any one speaker from a set of 1000 wanted individuals is engaged. His reference sample may be in Pashtu or Farsi, but the intercepted call may be in Arabic or English. For obvious reasons, the impact of the cross-language problem on these systems remains undisclosed, but neither has it received much attention in published research on auditory or automatic speaker recognition. This is all the more surprising since probably the majority of countries have become, or have always been, multi-ethnic and/or multi-lingual. On the auditory phonetic and linguistic side Bahr and Frisch (2002) reported that the speaker-discrimination performance of monolingual (American English) listeners decreased for cross-language stimulus pairs (English–Spanish), corroborating earlier studies by Goggin, Thompson, Strube and Simental (1991) and Hollien, Majewski and Doherty (1982). On the other hand, in a recent study that combined 'technical' factors of landline and cell-phone transmission with speaker-related features such as dialect and language, Betancourt and Bahr (2010) found that the language factor (English–English vs English–Spanish) had the least influence on the responses of their listeners.[1] More research is therefore needed. At any rate, §6 of the Code of Practice of the IAFPA (IAFPA 2004) issues a general warning to practitioners against carrying out speaker identification 'in languages of which they are not native speakers', and 'Members should exercise particular caution if the samples for comparison are in different languages.' The

latter clause clearly reflects the cross-language problem. In the terms of the quoted case, the reason for such caution is evident: Even in the unlikely event that the forensic expert is able to speak Igbo, the different languages involved pose a key problem for 'conventional' acoustic-phonetic-linguistic speaker recognition, since at least all *language*-related features have to be excluded from the comparison. For instance, it would be all but impossible to detect a difference in terms of regional dialect that, taken on its own, must lead to an exclusion of identity, even if all voice-related parameters or tempo, rhythm etc. may exhibit substantial similarities.

A small number of studies with automatic speaker-recognition systems have investigated certain aspects of the cross-language problem. In an early study based on the English and Dutch portions of the multilingual NATO N4 corpus, Zissman, van Buuren, Grieco, Reynolds, Steeneken and Huggins (2001) investigated two variables which they called 'cross-style' (read/spontaneous speech) and 'cross-language' training. Unfortunately, since both variables were combined it is impossible to determine their individual effects. The experimental setup consisted of read English and Dutch speech (the 'North Wind' text) produced by 30 Dutch L1 speakers as training material and short 'tactical' speech in English as test material, mostly the names for the characters according to the NATO alphabet. Three speaker-recognition systems were used for the closed-set identification tests. One produced the same high EER for both the English-only and the Dutch–English condition (33%). The second system was tested with two frequency ranges. In the 4 kHz mode the cross-language EER went down from 21% to 18% whereas it rose from 8% to 12 % in the 8 kHz mode. The third system produced 8% and 9% EER for the same-language and cross-language conditions. These results lead the authors to conclude that 'the impact of the training/testing language mismatch is system dependent' (Zissman et al. 2001: 2–5). In an extensive study evaluating 12 international speaker-recognition systems of the time (c. 2003), van Leeuwen and Bouten (2004) investigated, together with a number of other parameters, the systems' performance in two cross-language tasks. Considering that the complete speech material was taken from telephone conversations of real cases, the study is of special relevance to the forensic environment, although today the general level of performance has improved greatly. In one experiment speaker models were trained in Dutch, and test audios were chosen from other languages ('mostly English'; 2004: 77) and in another experiment the languages of models and tests were reversed. Compared to experiments with monolingual Dutch and English data, the cross-language data caused EERs to rise by 6.2% (Dutch models) and 9.0% (Dutch tests).[2] Lu, Dong, Zhao, Liu and Wang (2009) hypothesised that the 'language gap', i.e. the reduction in performance of many speaker-recognition

systems, may be due mainly to the fact that these systems were developed on the basis of English data. Using a JFA (Joint Factor Analysis) approach in which a high-dimensional GMM supervector is typically split into a speaker-related and a session-related subspace, they introduced an additional language subspace, modelled by multilingual data from two corpora with a total of 18 languages and 62 hours of speech, to provide what they called Language Factor Compensation (2009: 4218). The tests showed that for male and female subjects EERs decreased after applying language compensation. As could be expected, the effect on English-only samples was small (5.13% before and 4.99% after compensation for males, 7.84% and 7.11% for females), but larger for non-English samples (males: 8.92%/8.09%; females: 11.42%/9.8%). The fact that EERs for female subjects were in general notably higher than for males was not discussed by Lu et al., but shows an interesting parallel to the findings of the present and earlier investigations (see below). Incidentally, the system described by Lu et al. was among those participating in the NIST speaker-recognition evaluation of 2006. Although the cross-language problem was part of the evaluation plan, the summary report by Przybocki, Martin and Le (2007) treats it only marginally, discussing the performance of one (unquoted) system that was said to be typical of most participating systems. The tests were based on a part of the Mixer corpus that contains mainly Arabic, Russian, Chinese (Mandarin) and Spanish speech samples in addition to English (Cieri, Campbell, Nakasone, Miller and Walker 2004: 628; Campbell, Nakasone, Cieri, Miller, Walker, Martin and Przybocki 2004: 31). The most general finding was 'that performance is clearly superior for the matched (same-language) trials than for the unmatched' (2007: 1957), and this applies to English–English as well as the respective non-English–non-English samples. According to the respective DET curves (2007: 1957, Fig. 9), the average difference is about 5%. The language aspect was then analysed separately for target and non-target trials. It was observed that same-language target trials produced superior results compared to different-language target trials, but the effect was much clearer for the (same-language) non-English data. On the other hand, when the non-target trials were split for the language effect, the same-language non-English results were clearly worse than the English–English results. The authors do not provide a full explanation for this pattern but speculate that 'for non-English data … the system may be doing language recognition as much as it is doing speaker recognition' (2007: 1957). Irrespective of all methodological and procedural differences between the studies discussed above, it would seem that the capacity of the normalisation procedure may be the key to the cross-language problem. A good example is the dissertation by Bautista Tapas (2005: chs. 2–8), where different normalisation algorithms are investigated with

respect to factors such as type of channel, length of speech samples, number of training sessions, speaking style and speed, speaker sex, and also language.

At present one can say that automatic forensic speaker-recognition systems may well be affected by the cross-language problem, yet, unlike the phonetic-acoustic method, not in a principal but rather in a quantitative way. The main reason is that advanced automatic systems do not use highly language-specific 'high-level' features such as dialect, sociolect, intonation patterns, phonetic and linguistic parameters of hesitations etc. but 'low-level' acoustic features, mostly sets of cepstral coefficients, that are characteristic of the general resonance behaviour of the vocal tract of a speaker, and thus much less of the language involved – provided that a sufficient amount and quality of speech material is available (Gonzalez-Rodriguez, Fierrez-Aguilar and Ortega-García 2003; Gonzalez-Rodriguez, Ramos-Castro, García-Gomar and Ortega-García 2004; Gonzalez-Rodriguez, Drygajlo, Ramos-Castro, García-Gomar and Ortega-García 2006; Drygajlo 2007; Ramos-Castro 2007; Przybocki et al. 2007; Sturim, Campbell, Dehak, Karam, McCree, Reynolds, Richardson, Torres-Carrasquillo and Shum 2011). Current commercially available systems for FASR[3] typically rely on components such as universal background models (UBMs) and reference populations (also termed 'normalisation cohorts') that are constructed to match the conditions of each individual case as closely as possible, including not only technical characteristics of the transmission channel, but also speaker-related features such as sex, speaking mode, and particularly language.

In principle, several types of language mismatch are possible, the most extreme one implying different languages of speaker models, test samples and reference populations. Analogous to the 'technical' sources of channel mismatch, such a setting would have to deteriorate the performance of a system (Agnitio 2009: 90). The present investigation is based on the typical forensic situation described above, i.e. with a suspect speaker's sample in language A and a test sample in language B, which creates the mismatch. It is also assumed that a reference population matching the language of the speaker model, A, is available. This fact has to be kept in mind when comparing the results with those of the other studies. Compared to a setting without mismatch, i.e. with speaker model, test sample and reference population in the same language A, the increased 'dissimilarity' between a test sample in language B and a reference population in language A (the speaker model still being in language A) could lead to lower likelihood scores, i.e. reduce false acceptances – in principle at the cost of increasing false rejections. In principle, matching reference population and test sample for language is also possible but will affect the response of the system in the opposite way: if the test samples are made more 'similar' to the reference population than to the speaker model, the amount of false rejections

will be reduced but at the same time the number of false acceptances will increase, which is unacceptable from a forensic point of view. An example for this will be provided in the Discussion.

Experiment

Subjects

In forensic cases such as the one mentioned above, the vast majority of multilingual speakers are not perfectly bilingual but speak one language better than another. The traditional terms 'first' and 'second language' (L1, L2) have to be applied in a broader sense here because they imply a chronological order that actually may not exist. For example, unlike their parents, children of first-generation Turkish immigrants in Germany usually learn both Turkish and German simultaneously, and interference phenomena of vocabulary, grammar and pronunciation go in either direction. For the purposes of this investigation, a total of 75 subjects were selected and divided into six language groups. Groups 1 to 4 consisted of nearly equal numbers of a) native speakers of German who had learned Russian (n = 10), Polish (n = 15), English (n = 12) or Spanish (n = 17) as L2 or even L3 at school and/or at university, b) native speakers of these languages, who had learned German as L2 or L3 at school (exchange students to Marburg University) and c) true bilinguals, mostly children of Russian, Polish and Spanish parents, who grew up in Germany. Group 5 consisted of 10 Chinese exchange students from Peking, who had German language and linguistics as a major and spoke the language fluently. Group 6 consisted of 11 linguistics students at the University of Barcelona, who were perfectly bilingual in Spanish and Catalan.[4] Here, Spanish was arbitrarily declared L1.

All subjects were female. A rather trivial reason is that it turned out impossible to find enough male students to balance each of the six groups for sex. However, from a scientific point this problem made it possible to make a virtue out of necessity, since it has been found earlier that female voices constitute a greater challenge to automatic speaker identification than male voices. As was mentioned above, higher EERs for female subjects were also reported by Lu et al. (2009). Künzel (2010: 269f.) attributed significantly higher EERs in the identification/distinction of monozygotic female twins as compared to male twins to their higher F0 and the concomitant broader spacing of harmonics, which results in less (dense) spectral information, a fact that also explains why formant centre frequencies and bandwidths of female voices are in general more difficult to measure (Peterson and Barney 1952: 181; see also House 1959).

Speech material

The speech material consisted of 90 to 120 seconds of read and spontaneous speech. Subjects were asked to read the 'North Wind and the Sun' text and to describe 'life as a student at Marburg' spontaneously. One half of the speakers started these tasks in German and then continued in the respective foreign language, and the order was reversed for the second half. An analogous procedure was used for the Spanish–Catalan group. Here, the Spanish version of the 'North Wind' was read. Since a Catalan version was not available a similar fable text was used ('La llebre i la tortuga', 'The Hare and the Tortoise'). For reasons of organisation it was not possible to keep the technical recording conditions constant for all groups of subjects, which introduced an unwanted source of variability to the acoustic data. Most speakers were recorded in a studio cabin at Marburg University Phonetics Laboratory, but some had to be recorded at their respective homes, which involved features like room reverberation and extraneous noise. The Spanish–Catalan bilinguals were recorded in their respective homes or a quiet office room in Barcelona. However, all parts of the recording of each speaker were made at the same location. All recordings were made at 44.1 kHz/16 bits. Non-speech events such as clearing of throat, cough, laughter etc. were removed from the speech signal using *Adobe Audition*.

Procedure

In order to create telephone and Skype transmission channel characteristics the cleaned directly recorded audio files of all subjects were subjected to the following treatment:

1. Playback through a loudspeaker with a linear frequency response into the microphone of a standard Alcatel 4020 landline office telephone set, transmission as a local call to another standard telephone set with an attached professional galvanic interception device, and re-recording on computer through the analog input of the sound card.[5]
2. Playback through a loudspeaker with a linear frequency response into the microphone of either a Sony Ericsson W890 or LG GD510 cell phone, GSM transmission via the O2 network inside Germany; re-recording performed as described in 1.
3. (For the Spanish and Catalan material only) feeding the audio files directly into the Skype program (i.e. bypassing the internal loudspeaker) for coding and transmission via Internet, and re-recording on another computer using the same procedure in reverse.

The automatic FASR system Batvox 3.1 (see the technical specifications in Agnitio 2009: 107f.) processes only files of the format Windows WAV at 8 kHz/16 bits. Audio data were converted accordingly using the AVS audio converter software. Normalisation for channel, gender, type of speech, language, etc. is imperative in cases 'in which the homogeneous quality of the data cannot be guaranteed' (Agnitio 2009: 16f., 90f.). For the purpose of this experiment, it was decided to use the 'Identification Mode' of operation. Results are delivered as normalised scores that can easily be integrated into error matrices in order to calculate FA and FR rates, and, finally, EERs. For all speakers of the first five groups, the fable text plus part of the spontaneous text (to make the overall duration one minute) in German were used to train the respective speaker models (targets). The German and the Russian/Polish/English/Spanish/Chinese samples were used as test audio. For each of the 1 + 2 (direct + telephone, GSM) transmission channel characteristics (German–Spanish: 1 + 3: direct + telephone, GSM, Skype) separate reference populations were used. At this juncture it is important to note that, although the language of the speakers of all four reference populations was the same, i.e. German, the speakers themselves were not identical, and their total numbers were also different, since they had been collected over time on the basis of previous forensic case material and/or data from research projects that happened to exhibit the respective channel characteristics. For tests with group 6 that contained the Spanish–Catalan bilinguals, the first 60 seconds of the Spanish material were used to train the speaker models, and the Spanish and Catalan samples served as test samples. Different parts of the GAUDI database (kindly provided by the Spanish Guardia Civil) served as reference populations.

A special feature of the normalisation procedure of Batvox is the option to use so-called 'case impostors'. The term is used quite differently from the traditional meaning of 'impostor' as 'non-target speaker' and denotes a set of speakers who are definitely not identical with the speaker under test but exhibit some similarities, primarily in terms of channel characteristics, and in this case language. Thus the system may recognise certain acoustic resemblances as irrelevant and reduce *a priori* the probability for false acceptance errors. Technically, the impostors are used as a Z-norm cohort in what may be called a second normalisation process, after application of the T-norm, and serve to reduce the misalignment in the event that the available T-norm cohort is less-than-perfect. Since the number of impostors is usually small (between 3 and 10) the second normalisation is based only on the mean of the cohort scores but not on its variance.[6] The option is particularly useful in cases where some speakers are available who can be excluded as the suspect speaker, for instance the very person(s) the suspect speaker is talking to. In the present experiment,

three to five female subjects reading the same texts in German, with the respective L1 backgrounds and the same transmission characteristics as the test speakers, were used as case impostors. For the Spanish–Catalan experiment, case impostors consisted of five Spanish-speaking female subjects recorded in the same channel conditions.

The following tests were carried out for each language group: every speaker's German (groups 1–5) or Spanish (group 6) speech sample was used to train the speaker model. In the cross-language task it was to be compared with the same speaker's L2 speech sample, which would be a match (target trial), and with all other speakers' L2 samples, which would be no-matches (impostor trials). In the same-language task, i.e. with a German speaker model *and* a German test file (for group 6: Spanish speaker model and Spanish test file) only impostor trials (no-matches) were used.[7] Here is an example: The German–Polish group consists of 15 speakers. The number of cross-language comparisons is 15 matches + 15 × 14 no-matches = 225; the number for same-language comparisons is 15 × 14 = 210 no-matches. For the purpose of this study and its small number of speakers per language group, it was considered adequate to use equal error graphs to represent the performance of the system. EERs were calculated using the Biometrics 1.2 software (Biometrics 1.2, 2012).

Results

As was set out in the Introduction, the current study is based on, and limited to, the typical forensic scenario where the speaker model is in language A, the test sample in language B and a reference population matching the speaker model in terms of language A is available. This special setting does not require the calculation of same-language false rejection scores ('FRsame'). In the present case, reference population and speaker model are both either in German whilst the test sample is in one of five other languages, or in Spanish, with the test sample in Catalan. In all test sessions three error functions were calculated on the basis of the likelihood scores:

1. FRcross: cross-language false rejections of a match, e.g. speaker model = German, test sample = Polish (cross-language target trials);
2. FAsame: same-language false acceptances of a no-match, e.g. speaker model *and* test sample = German (same-language impostor trials);
3. FAcross: cross-language false acceptances of a no-match, e.g. speaker model = German, test sample = Polish (cross-language impostor trials).

The scale for the likelihood scores on the abscissa is given by the system. Figure 1a shows a typical example of the three functions for the German–Polish

bilinguals in the landline telephone transmission condition. EER values for FRcross/FAsame and for FRcross/FAcross, are 0.5%, and 0%, respectively. At first glance the two distributions for false acceptances are quite close to each other and essentially parallel, with the distribution for FAcross (dotted line) shifted to the left by a small margin and thus slightly more separated from the distribution of FAsame. Quite obviously, the EER is hardly influenced by the fact whether the languages of the impostors are the same or different, and this picture is typical of the majority of language and transmission conditions. Figure 1b shows a scattergram of the raw data (LR scores) for FAsame and FRcross which were used in Figure 1a. There is only one case of overlap (arrow) of FR (triangles) and FA (dots) values. The highest EERs of all tests, i.e. the relatively worst performance of the automatic system, were obtained for the German–Spanish voice data in the GSM condition. As can be seen in Figure 2, the two distributions for false acceptances diverge with increasing LR scores. At the point of intersection of the (Spanish) matches (FRcross) the EERs are 5.9% for the same-language (German) impostors and 4.9 % for the cross-language (Spanish) impostors.

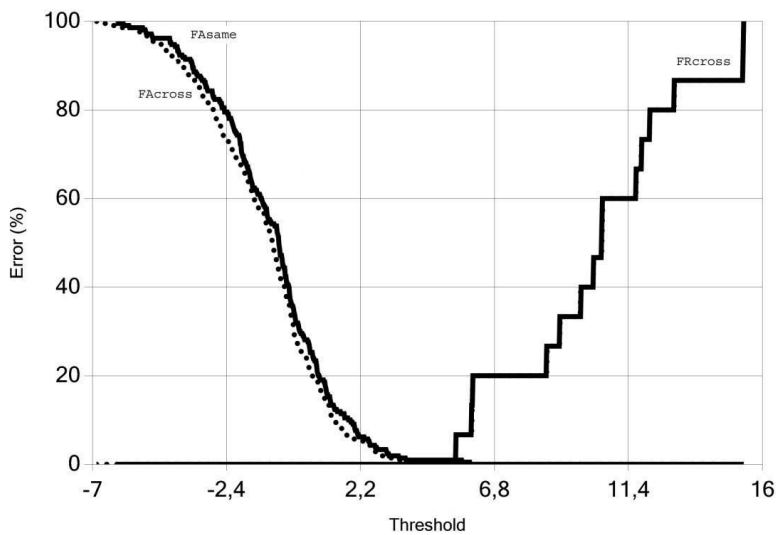


Figure 1a: Cumulative distributions of scores for same-language and cross-language comparisons of 15 female German–Polish speakers in the landline telephone transmission condition. The reference population selected for this test session consists of 63 female speakers of German. EERs are 0.5% and 0% for the same and cross-language condition.

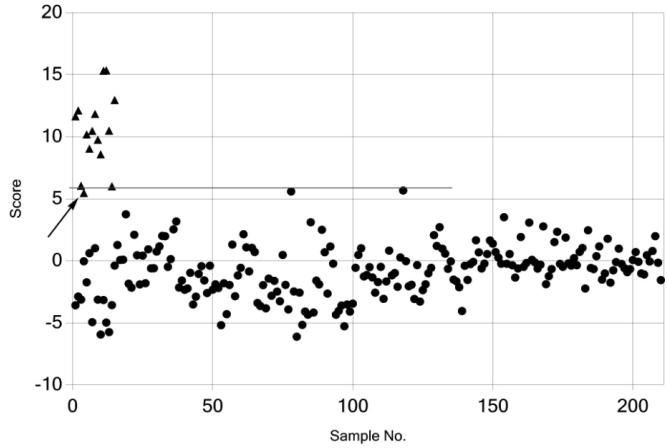


Figure 1b: Scattergram of scores for FRcross (triangles) and same-language impostors (dots) used in Figure 1a. The only case of overlap is indicated by an arrow.

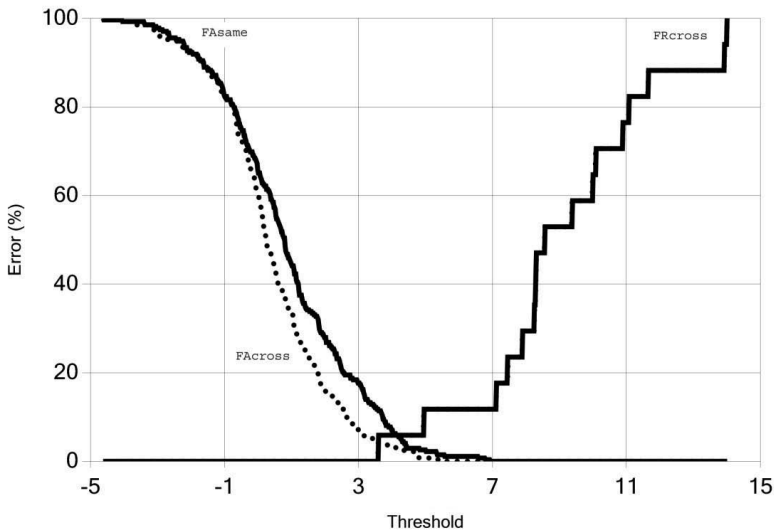


Figure 2: Cumulative distributions of scores for same-language and cross-language comparisons of 17 female German–Spanish speakers in the GSM transmission condition. The reference population consists of 74 female voices. EERs are 5.9% and 4.9% for the same- and cross-language condition.

Table 1: Equal error rates (per cent) for bilingual speakers for four types of transmission channel characteristics; 'same' (grey background) and 'cross' refer to the languages of speaker models/impostors.

Bilingual group	No. of speakers	Reference population	Speaker model	Test audio	Impostors	Direct rec.	Landline tel.	GSM tel.	VoIP (Skype)
German–Russian	10	GER	GER	RUS	GER	0	0	0.6	
		GER	GER	RUS	RUS	0	0	0	
German–Polish	15	GER	GER	POL	GER	0	0.5	0.2	
		GER	GER	POL	POL	0	0	0	
German–English	12	GER	GER	ENG	GER	0.8	0.4	0.4	
		GER	GER	ENG	ENG	0.4	0.4	0.8	
German–Chinese	10	GER	GER	CHI	GER	0.6	0	0	
		GER	GER	CHI	CHI	0	0	0.6	
German–Spanish	17	GER	GER	SPA	GER	0.2	1.1	5.9	1.5
		GER	GER	SPA	SPA	0.2	0.4	5.0	0.2
Spanish–Catalan	11	SPA	SPA	CAT	SPA	0	0.9	0	0
		SPA	SPA	CAT	CAT	0	0	0	0
				Grand mean		**0.2**	**0.3**	**1.1**	**0.4**
					same	0.2	0.5	1.2	0.7
					cross	0.1	0.1	1.0	0.1
					difference	0.2	0.4	0.1	0.6

Table 1 contains EERs for all six bilingual groups and transmission conditions. It can be observed that the general level of EERs is quite low, with 36 out of the 40 values below 1%. The remaining four values all pertain to the same speaker group (GER–SPA): two between 1% and 2% (landline telephone, Skype) and the remaining two between 5% and 6% (GSM; cf. Figure 2). Comparing the percentages for same-language (grey shaded lines) and cross-language EERs of each bilingual group it becomes obvious that for 9 of the 20 (vertical) pairs of data the cross-language condition involves slightly lower EERs than the corresponding same-language condition, whereas the opposite relation occurs in two cases only, GER–ENG (GSM), GER–CHI (GSM). For the remaining nine cases, both values are identical, seven of them 0%. The differences between pairs of EERs in the same-language and different-language

tests vary between 0.1% and 0.6% for the six speaker groups in the four transmission characteristics. The means for all same- and cross-language EERs are 0.6% and 0.4%. Figure 3 shows the means for all speaker groups and transmission conditions at a glance. All same-language (grey) columns are larger than the neighbouring cross-language (black) columns. The figure also shows that EERs are lowest for the direct-recording condition, which may be considered here as a benchmark, since it differs from the original studio-quality recordings only in terms of the reduced frequency range. Individual values are between 0% and 0.8% (cf. Table 1). As could be expected, EERs for landline telephone and cell phone recordings are generally higher, varying from 0% to 1.1% (landline) and 0% to 5.9%, respectively. Since the Skype transmission condition was available only for two of the six bilingual groups, the respective four EER values must be regarded with caution. At any rate they are within the variation for the two telephone conditions (0%–1.5%).

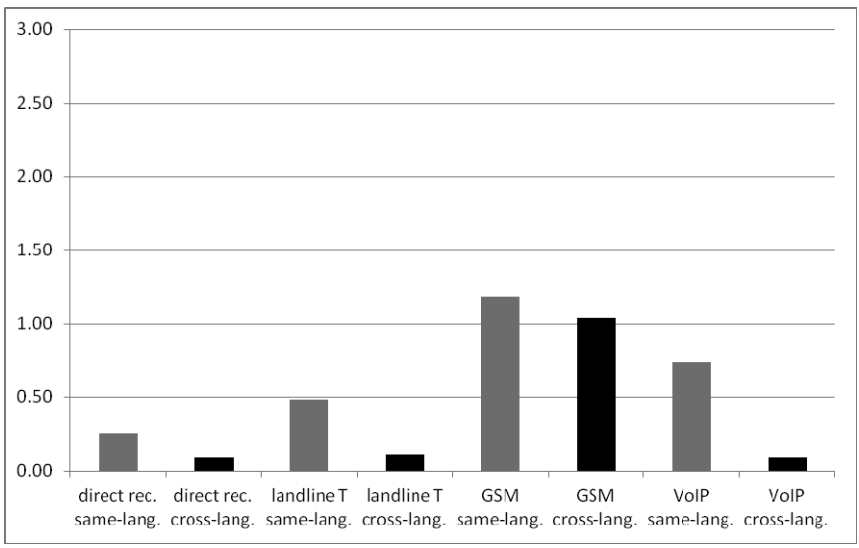


Figure 3: EERs (per cent) of same-language and cross-language comparisons for four transmission channels

The second main question of this investigation is about the magnitude of the cross-language effect in relation to the effect of the transmission channel. In order to investigate the question, two sets of data were derived from Table 1. The *between-languages* set contained the 20 differences between cross-language and same-language values (vertical pairs) of each group and transmission condition. The *between-channels* set consisted of 48 differences, three

each for the first eight horizontal lines and six for each of the last four lines. The data are displayed in Figure 4. All differences related to the cross-language condition and also the vast majority of channel-related differences are between 0% and less than 2%. A small number of channel-related differences extend to < 6%, caused by the relatively high EERs for the German–Spanish comparisons in the GSM condition. The class means of 0.35% and 0.8% are significantly different ($p = 0.04$, 2-sided t-test) which means that under the conditions of this experiment the magnitude of the cross-language effect was not larger, but in fact even smaller, than the effect of the landline telephone, GSM and Skype transmission as compared to the direct recording condition.

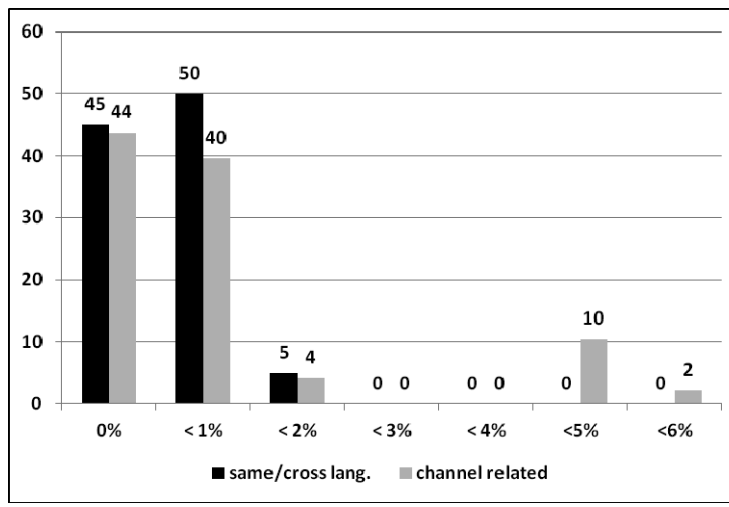


Figure 4: Relative distribution of EER differences related to the same- / cross-language condition, and to the four transmission channels

Discussion

Comparing the findings of the present study with those quoted above, the general level of performance in terms of EERs of the current system is superior. Several factors may have facilitated the recognition tasks: the amount of speakers in each language group was small (between 10 and 17), and there were obviously no extremely 'difficult' speakers, i.e. 'goats' and 'wolves' in the terminology of Doddington, Liggett, Martin, Przybocki and Reynolds (1998: 37). The speech material was homogeneous, since its major part, the fable text, was identical for all speakers inside the same-language group, and the small

portions of spontaneous speech consisted of descriptions of the same topic. Furthermore, speech samples in both languages of a speaker were recorded on the same day. For most speakers the recording conditions as well as the four transmission channels (including the telephone sets and mobile phones) were identical, and as far as Skype is concerned, the overall quality of transmission may have been atypically good (avoiding AD–DA conversion and bypassing the computer loudspeakers when sending and receiving the audio signal). The most important reason for the low EERs, however, is the performance of the automatic system itself, with its double normalisation procedure that includes the option to use *case impostors*, a feature that is specifically useful for cross-language recognition tasks.

According to previous studies, cross-language speech samples tend to degrade the performance of automatic speaker-recognition systems. However, without knowing the details of the architecture of these systems, it is impossible to identify the immediate cause(s). With respect to their own system, Lu et al. (2009: 4217) argue, 'the main reason may lie in the fact that the whole system is built up mainly based on English development data'. Discussing the performance of one (unnamed) system, Przybocki et al. (2007: 1957) suggest, 'non-English conversations receiving less evaluation emphasis' – whatever that may mean – as a possible cause for clear differences between cross- and same-language speech. Van Leeuwen and Bouten (2004: 80) presume that most of the speaker-recognition systems they had tested 'have been developed with English data (e.g. for estimating a universal background GMM model), and that if there would be any language effect, it would be advantageous for English'. On the other hand, taking account of the mismatch and trying to compensate it can actually improve the performance (Lu et al. 2009). In other words, it seems that the process of normalisation – irrespective of its system-specific details – may play a key role. Another reason for the relatively small size of the cross-language effect is probably the fact that, unlike the systems discussed earlier, the present system was designed *ab initio* without special regard to the English language. For instance, the two UBMs for male and female speakers each consist of more than a thousand voices of many languages.

In the present investigation, the overall performance of the automatic system for cross-language voice comparisons was equal to or at times slightly better than for same-language comparisons. Several reasons may have contributed to this finding. The special setting of parameters tailored to the typical forensic situation described above produces a language mismatch only between speaker model and test sample, whereas a reference population and also a number of 'case impostors' are available that match the speaker model in terms of

language. This constellation of parameters provides the possibility of what may be called *optimal normalisation* of the likelihood scores.[8] A consequence of the forensic setting is that, unlike previous studies which used impostor trials and target trials in both cross-language and same-language recognition tasks, only impostor trials (FAsame) but no target trials were carried out for the same-language condition. If it is assumed that in principle – other parameters such as reference population and case impostors being equal – same-language recognition tasks will produce overall higher scores than different-language tasks, then this would result in a right-shift of *all* scores (FAsame, and FRsame if available) in comparison with the cross-language task, and thus the overlap between FAsame and FRcross (EER) will also be larger than between FAcross and FRcross.

The importance of the normalisation procedure can be demonstrated *ex negativo* by the following example taken from the German–Chinese data. Let us reverse the 'canonical' setting used hitherto and use the Chinese rather than the German speech samples to train the speaker models, and the German samples as tests. The German reference population shall be left unchanged. The new setting precludes an optimal normalisation since no Chinese reference population is available. Still using the German reference population causes what may be called *non-optimal normalisation*. With the language mismatch now between reference population and speaker model the performance of the system degrades. Table 2 contains the EERs obtained with optimal and non-optimal normalisation. It can be seen that both same-language and cross-language comparisons involve higher EERs when the normalisation is not optimal. Differences are considerable for the direct and telephone conditions. In the GSM condition the increase is only small in the same-language condition. The changes can be observed in a more detailed way in Figure 5. It shows FA and FR obtained with optimal normalisation on the left side (Figure 5a1, 2, 3). The three graphs should be compared to those on the right side that are based on the results with non-optimal normalisation (Figure 5b1, 2, 3). In the latter graphs the scores for FAcross have decreased, i.e. the distribution has shifted to the left but at the same time the distribution of FRcross has shifted to the left even more. Since FAsame scores have remained largely unchanged, the largest overlaps are now between FRcross and FAsame. In principle, this result is the same that was obtained for the optimal normalisation, except for the fact that the absolute level of EERs is higher. Put differently: without optimal normalisation the performance of the system will decrease considerably. This result is in accordance with Bautista Tapias, who used an earlier version of the present system (2005: 297f.).

Table 2: Equal error rates (per cent) for the German–Chinese speaker group in same- and cross-language comparisons with and without optimal normalisation.

normalisation	reference population	speaker model	test audio	impostors / (model)	language	direct rec.	landline tel.	GSM tel.
optimal	GER	GER	CHI	GER (same)	same	0.6	0	0
non-optimal	GER	CHI	GER	CHI (same)	same	8.9	10.5	1.1
optimal	GER	GER	CHI	CHI (cross)	cross	0	0	0.6
non-optimal	GER	CHI	GER	GER (cross)	cross	10	10	10

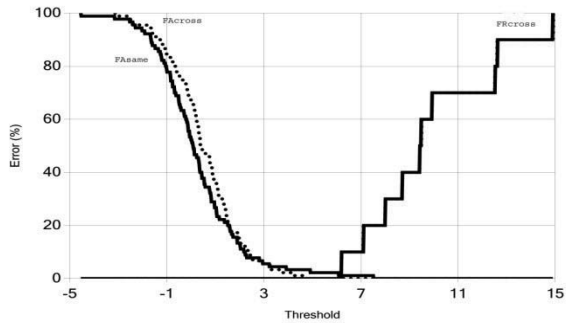
It would be of interest to modify the experimental setting in yet another way, by changing the language of the reference population, for instance, from German to Chinese, and observe the effects of optimal and non-optimal normalisations. This may be an issue in a forensic case where there is not enough German but enough Chinese speech material to calculate a speaker model. Then, the optimal normalisation procedure would consist in using a Chinese reference population for the Chinese speaker models and the German speech material as test sample.

At this juncture it should be reiterated that care must be taken when comparing different automatic systems, test plots and performance measures. Due to their individual architecture, systems may require different types of normalisation for the effects of language or transmission channel. The development of the system used here can serve as an example. As was stated earlier, the present version requires normalisation for channel, gender, type of speech, language, etc. in cases 'in which the homogeneous quality of the data cannot be guaranteed' (Agnitio 2009: 16f., 90f.), and this process is performed by the user selecting the optimal reference population. The forthcoming version will be based on the i-vector theory and its concept of *Total Variability Space* (Dehak, Dehak, Kenny, Brummer, Oellet and Dumouchel 2009). Here, the reference population will be separated from the normalisation cohort. While the expert user still has to select the former, the system automatically selects the latter out of an inventory of previously stored speaker models. One of the practical implications of this approach is that a speaker model and a test sample of certain minimum durations (40 and 7 seconds) are no longer required. Rather, the *total amount* of speech material must be 40 seconds or longer.

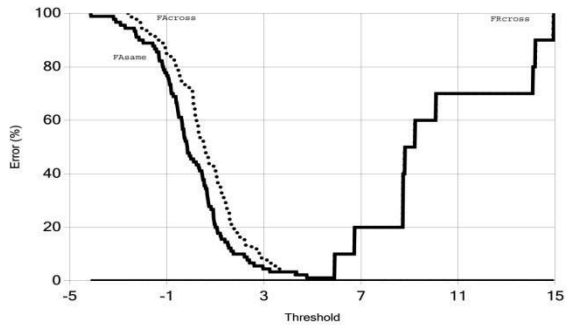
EERs should be used with caution for an assessment of the cross-language factor, since they reflect the performance of an automatic system only at one operating point. DET displays contain much more information but require more data than were available in the present experiment, with only 10 to 17 speakers per language group. Therefore it will be necessary to conduct cross-

5a) Reference population = German, Speaker model = German, Test = Chinese

1) direct recording



2) landline telephone



3) GSM telephone

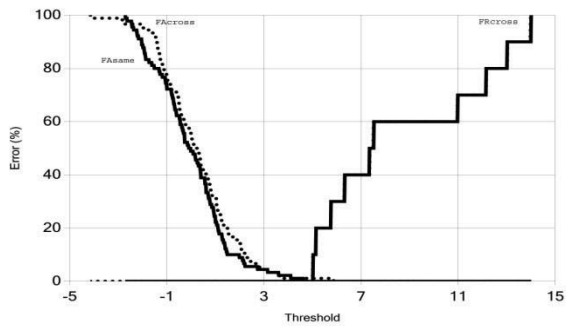
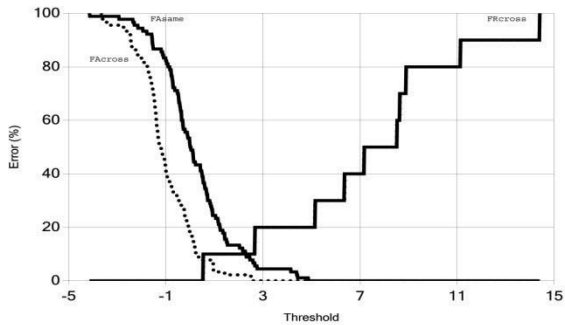


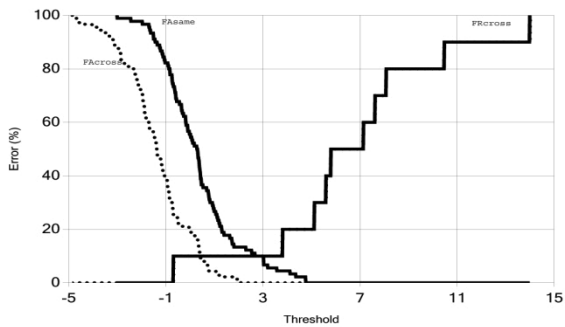
Figure 5: EER plots for same-language and cross-language comparisons of 10 female Chinese speaking German and Chinese in three transmission conditions, with and without language match of reference population and speaker model (abscissa = Score).

5b) Reference population = German, Speaker model = Chinese, Test = German

1) direct recording



2) landline telephone



3) GSM telephone

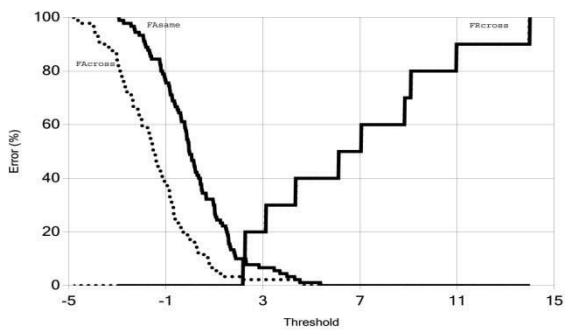


Figure 5 (cont.)

language experiments with much larger corpora. The deficit of EER as a criterion can be demonstrated with the three drawings on the left side of Figure 5. For the direct recording condition (Figure 5a1) the distribution of the FAsame scores is located slightly left to the FAcross distribution, i.e. most of its score values are smaller. However, at the point of intersection to the FRcross distribution, the tail of the FAsame distribution is higher, resulting in an EER of 0.6% whereas for FAcross/FRcross EER is zero (cf. Table 1). For the telephone condition (Figure 5a2) the picture is similar, yet here both EERs have the same values (0%). For the GSM condition (Figure 5a3), both FA distributions overlap most of the time, but here, against the general trend, the EER for the cross-language condition is slightly higher than for the same-language condition (0.6% vs 0%).

With respect to channel transmission characteristics, the results of the present study are in accordance with earlier results. Increases in EERs from high-quality direct recordings to landline and GSM recordings were to be expected due to the reduced amount of spectral information and data compression. Again, absolute values are rather low, which is probably the consequence of using well-adapted reference populations. Special caution is necessary with respect to the Skype condition. First, results are available from only two language groups so far. The more important reason is, however, that the transmissions were performed inside a small IP-network and are perhaps not realistic enough in forensic terms. However, first results from an ongoing investigation into the effect of VoIP transmission on automatic speaker recognition suggest that, all other conditions being equal, transmission inside Europe via Skype (from Germany to Romania) does not usually degrade the performance of the system more than does a standard GSM transmission. In general, it can be said that under the conditions of this experiment the impact of channel transmission characteristics on recognition results was generally larger than the language mismatch effect. Considering that female voices are under all circumstances more difficult to recognize than male voices (see also the results by Lu et al. 2009, Künzel 2010), it may be hypothesised that results for male subjects will be at least as good as the results of this study.

Conclusion

The experiment described in this article has shown – admittedly on quite a small database (75 speakers) – that using a normalisation procedure that accounts for different languages of speaker models and test samples can reduce this source of mismatch to a level that justifies the use of an (this!) automatic system in cross-language cases. In fact, the false-acceptance probability for

cross-language comparisons is not higher but may be even lower than for same-language comparisons. Due to the architecture of the system used here (in particular considering its option for a double normalisation), and also due to the forensic paradigm in which false acceptances are considered the worst kind of identification errors, it is mandatory to match reference population and speaker model in terms of language. If this condition is met, the remaining language mismatch in relation to the test sample can be neglected. This result holds not only for direct-recorded data but also for typical forensic channel conditions such as landline, GSM telephone and, with certain restrictions, Skype as one service of VoIP transmission. Here, coding algorithms used by other services should also be investigated – and in more realistic conditions, such as transcontinental data links at peak traffic hours. In essence, the claim that automatic speaker-identification systems of the kind described here are largely independent of language can be confirmed. However, since currently available systems differ greatly in terms of their architecture, all will have to be tested individually for cross-language effects.

About the author

Hermann J. Künzel is Professor of Phonetics at the University of Marburg, Germany. From 1985 to 1999 he was Head of the Speaker Identification & Tape Authentication Department of the Federal Criminal Police Office (BKA) in Wiesbaden, Germany. In the years 1980 to 1990 he was essential in the development of the acoustic-phonetic method of forensic speaker recognition (FSR) and has been working as a professional expert in FSR, speaker profiling, voice line-ups and non-speech related acoustic investigations (e.g. aircraft and shipping incidents) for courts and government institutions throughout Germany and worldwide. He has been applying automatic speaker identification to cases of lawful interception and in court since 2001.

Acknowledgments

Many thanks to Antonio Moreno (Agnitio, Madrid) for his technical advice, Anil Alexander and Oscar Forth (Oxford Wave Research) for helpful comments on the manuscript and for providing the BioMetrics 1.2 software that considerably facilitated drawing all figures containing error plots.

Notes

1 The authors hypothesise that this finding may be due partly to the test format (forced-choice discrimination rather than identification) and the

instructions that may have focused the listeners' attention to vocal rather than language-related features.

2 These figures have to be considered in the context that the lowest EER level of all systems was 12.1% for the test condition (60 s for training and 15 s for test audio).

3 E.g. LVIS by Loquendo (www.loquendo.com) or Batvox (www.agnitio.es).

4 The acoustic data were recorded by Claudia Schönfelder for her MA thesis 'Einfluss von Sprache und Übertragungskanal in der automatischen Sprecher-Identifizierung: Eine empirische Studie', Marburg 2010.

5 It goes without saying that the analog 'detour' of recording telephone speech is not typical of the interception process used normally by German police. However, it is still used in a number of cases. At any rate the frequency response of the galvanic device and the noise produced by the AD conversion create degradations of the speech signal and can thus introduce another source of mismatch if (only) a part of the material was produced this way. Playback for the landline and GSM transmissions was performed in a sound-treated cabin.

6 In fact, the system calculates the average of the case impostor scores and subtracts it from the score obtained for the comparison of the suspect speaker model and the test sample (Agnitio 2009: 18).

7 This is due to the fact that only one speech sample per speaker per language had been recorded. Although the total duration of the samples (2 minutes) would have provided more than enough material for creating a second same-language sample, this was not considered useful because it was all recorded in one session. Thus intra-speaker variability would have been too small, i.e. restricted to the type of text (read–spontaneous).

8 In fact, this setting is recommended in the Manual (Agnitio 2009: 94). It also contains an example with German and Spanish speech data. The DET plot on p. 93 shows that matching reference population and speaker model for language increases the performance 'mainly in the zone of false acceptance'. See also Bautista Tapias 2005: 200.

References

Agnitio (2009) Batvox 3.0 Basic User Manual. Madrid.

Bahr, R.H. and Frisch, S. (2002) The problem of code switching in voice identification. In A. Braun, and H. Masthoff (eds) *Phonetics and its Applications: Festschrift for Jens-Peter Koester on the Occasion of his 60th birthday* 86–96. Stuttgart: Steiner.

Bautista Tapas, R. (2005) Sistemas forenses de reconocimiento automático de locutores. Determinación y análisis de sus variables más críticas. Proyecto fin de carrera, Universidad Politécnica de Madrid.

Betancourt, K.S. and Bahr, R.H. (2010) The influence of signal complexity on speaker identification. *International Journal of Speech, Language and the Law* 17(2): 179–200.

Biometrics 1.2 (2012) Performance metrics software user guide. Oxford Wave Research Ltd (www.oxfordwaveresearch.com).

Campbell, J.P., Nakasone, H., Cieri, C., Miller, D., Walker, K., Martin, A.F. and Przybocki, M.A. (2004) The MMSR bilingual and crosschannel corpora for speaker recognition research and evaluation. *Proceedings of Odyssey 04 Speaker and Language Recognition Workshop, Toledo (Spain):* 29–32.

Cieri, C., Campbell, J.P., Nakasone, H., Miller, D. and Walker, K. (2004) The Mixer corpus of multilingual, multichannel speaker recognition data. *Proceedings Information for the Defense Community, DTIC Conference Paper*: 627–630.

Dehak, N., Dehak, R., Kenny, P., Brummer, N., Oellet, P. and Dumouchel, P. (2009) Support vector machines versus fast scoring in the long-dimensional total variability space for speaker verification. *Proceedings ISCA Interspeech 2009 Brighton,UK*: 1559–1562.

Drygajlo, A. (2007) Forensic automatic speaker recognition. *IEEE Signal Processing Magazine* 24: 132–135. http://dx.doi.org/10.1109/MSP.2007.323278

Doddington, G., Liggett, W., Martin, A., Przybocki, M. and Reynolds, D.A. (1998) Sheeps, goats, lambs and wolves: a statistical analysis of speaker performance in the NIST 1998 speaker recognition evaluation. *Proceedings of International Cenference on Spoken Language Processing*, Sydney: 37–40.

Goggin, J.P., Thompson, C.P., Strube, G and Simental, L.R. (1991) The role of language familiarity in voice identification. *Memory and Cognition* 19: 448–458. http://dx.doi.org/10.3758/BF03199567

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.

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. http://dx.doi.org/10.1016/j.csl.2005.08.005

Hollien, H., Majewski, W. and Doherty, E.T. (1982) Perceptual identification of voices under normal, stress and disguise speaking conditions. *Journal of Phonetics* 10: 139–148.

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

IAFPA (International Association for Forensic Phonetics and Acoustics) (2004) Code of Practise. www.iafpa.net/code.htm.

Künzel, H.J. (2010) Automatic speaker recognition of identical twins. *International Journal of Speech, Language and the Law* 17: 251–277.

Lu, L., Dong, Y., Zhao, X, Liu, J. and Wang, H. (2009), The effect of language factors for robust speaker recognition. *IEEE – ICASSP 2009*: 4217–4220.

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. http://dx.doi.org/10.1121/1.1906875

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. http://dx.doi.org/10.1109/TASL.2007.902489

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

Sturim, D., Campbell, W., Dehak, N., Karam, Z., McCree, A., Reynolds, D., Richardson, F., Torres-Carrasquillo, P. and Shum, S. (2011) The MIT LL 2010 speaker recognition evaluation system: scalable language-dependent speaker recognition. *IEEE – ICASSP 2011*: 5272–5275.

van Leeuwen, D. and Bouten, J.S. (2004) Results of the 2003 NFI-TNO forensic speaker recognition evaluation. *Proceedings of Odyssey 04 Speaker and Language Recognition Workshop, Toledo (Spain)*: 75–82.

Zissman, M.A., van Buuren, R.A., Grieco, J.J., Reynolds, D.A., Steeneken, H.J.M. and Huggins, M.C.(2001) Preliminary speaker recognition experiments on the NATO N4 corpus. *Proceedings RTO IST Workshop on Multilingual Speech and Language Processing, Aalborg, Denmark, (RTO-MP-066)*: 2.1–2.6.