The WCL-1 System in the 2003 NIST Speaker Recognition Evaluation and 2003 NFI/TNO Forensic Speaker Recognition Evaluation

Todor Ganchev, Nikos Fakotakis, George Kokkinakis

Wire Communications Laboratory, Dept. of Electrical and Computer Engineering,
University of Patras, GR-26500 Rion-Patras, Greece
tganchev@wcl.ee.upatras.gr

Abstract

In the present work we discuss the results, which our speaker verification system, WCL-1, obtained in the 2003 NFI/TNO Forensic Speaker Recognition Evaluation. These results, together with the ones obtained in the 2003 NIST Speaker Recognition Evaluation, give opportunity for in depth analysis of the various aspects of real-world application of the speaker recognition technology. Based on the detailed analysis of the speaker verification performance obtained in the different subtasks, we identify the virtues and disadvantages of the WCL-1 system and its potential areas of use.

1. Introduction

During the last ten years the speaker recognition community has enjoyed a significant acquirement – annual technology evaluation forums, referred to as Speaker Recognition Evaluation (SRE) campaigns, where the practical significance of the different approaches can be corroborated in a common experimental setup. Similarly to many other areas, the worldwide collaboration and free exchange of ideas led to technology boost in the speaker recognition field as well.

The SRE campaigns are organized as technology evaluation forums, where although there are multiple different evaluation tasks, no specific application is named. Instead, the SREs aim at evaluation of basic technology in "extreme" conditions, which reflect any potential variability of the real-world environment. For each SRE campaign, a specific database is designed. This database utilizes real-world recordings, which are post-processed to facilitate the experimentations. Special care is paid for securing controlled conditions in the various subtasks. These subtask target at evaluation the influence of specific aspects of variability that can be observed in real-world environment. For instance, mismatch between train and test conditions due to different transmission standards, telephone handsets, environment, speech duration, etc.

Due to the different focus and different degree of difficulty of consecutive SREs, it is attractive to evaluate how a given system performed in two or more evaluations. Comparison between the 2004 NIST SRE [1] and 2003 NFI/TNO Forensic SRE [2] for four different systems is available in [3]. In contrast to [3], in the present work we consider the NIST SRE and NFI/TNO Forensic SRE campaigns, which took place in year 2003. Since these evaluations were held within six months one from another, we had the chance to use the same system, which is referred here to as WCL-1 in both of them. These two evaluation campaigns are of statistically different difficulty and represent different potential applications. Therefore, comparative analysis of the results would be of particular importance, when the appropriateness of a system for a given class of real-world applications is evaluated.

2. The WCL-1 system

The WCL-1 system participated in several SRE campaigns. In the present section, we describe its configuration, as it was during the 2003 NIST SRE and 2003 NFI/TNO Forensic SRE campaigns.

2.1. Architecture of the WCL-1 system

In brief, the WCL-1 system is built on a modular structure, where for each target speaker a distinct expert is employed. Specifically, for each enrolled speaker, a personal probabilistic neural network (PNN) [4] is designed to recognize him/her among an unlimited number of other speakers. Both the client and the other speakers are represented by codebooks, and therefore, the speaker verification process is dealt with as to a typical two-class problem.

A simplified block diagram of the PNN-based speaker verification system, WCL-1, is presented in Figure 1. The upper part of the figure summarizes the process of training, where the process of building of the reference model, referred to as universal background codebook (UBgCB), as well as construction of the individual codebooks for the target speakers is shown. A personal PNN for each of the target users is created, by utilizing the reference codebook and the codebook created for the corresponding user. The lower part of the figure illustrates the operational mode of the system. The processing steps, WCL-1 performs for each test trial in order to make a final decision, are shown. In the following subsections, the main building blocks of our speaker verification system are described in details.

2.2. Speech pre-processing and speech features extraction

In both the 2003 NIST SRE and 2003 NFI/TNO Forensic SRE we deal with telephone quality speech, sampled at 8 kHz. Saturation by level is a common phenomenon for telephone speech signals. In order to reduce the spectral distortions it causes, a band-pass filtering of speech is performed as a first step of the feature extraction process. A fifth-order Butterworth filter with pass-band from 80 Hz to 3800 Hz is used for both training and testing. Then the speech signal is pre-emphasized with $a = 0.97$ and subsequently, windowed into frames of 40 ms duration, at a frame rate of 100 Hz using a Hamming window. Each frame is subjected to 1024-point short-time Discrete Fourier Transform, and then is passed through a set of 32 triangular band-pass filter-bank channels. We have accepted an approximation of the Mel-scale, with 13 linearly spaced filter-banks, lowest central frequency 200 Hz, highest 1000 Hz, and 19 log-spaced with highest central frequency 3690 Hz. Subsequently, 32 MFCC parameters are computed, after applying Discrete Cosine Transform to the

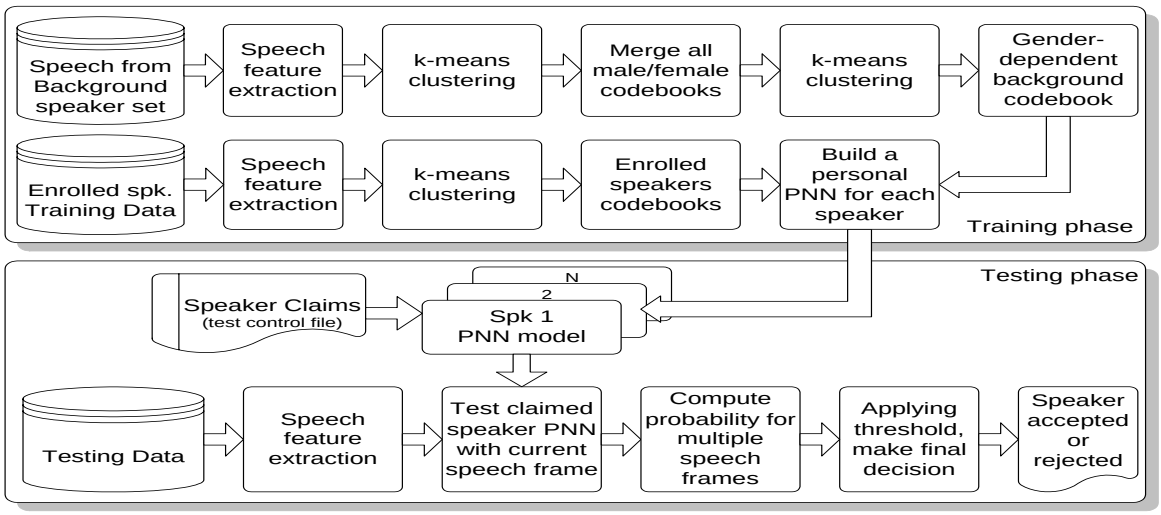

Figure 1: A simplified block diagram of the WCL-1 system

log-filter-bank outputs. This MFCC implementation is based on [5]. A comprehensive description of the MFCC computation steps is offered in [6], where various implementation strategies are evaluated. Only the feature vectors extracted for voiced speech frames are used to represent the speakers' identity. The voiced / unvoiced speech separation is performed by a modification of the autocorrelation method with clipping [7]. The unvoiced speech is more sensitive to noise and by this reason it is discarded. Finally, the WCL-1 system utilizes a feature vector composed of thirty-two Mel-frequency cepstral coefficients, logarithm of the fundamental frequency and the energy of speech frame. In fact, instead of the traditional ln(f0), we made use of ln(f0-$f0_{min}$), which we found out (see [8] for details) to be much more effective. The dynamic range of ln(f0-$f0_{min}$) is extended, which better corresponds to the relative importance of the fundamental frequency. The constant $f0_{min}$ = 55 Hz is selected, as 90% of the minimal fundamental frequency the pitch extractor can detect.

2.3. The client model design

The clients of the speaker recognition system are often referred to as target speakers. The speech recordings, collected during the enrolment session of a potential client are used for creating his personal model.

Because the complexity and computational demands of the PNNs depend strongly on the number and dimensionality of the training vectors, a k-means clustering algorithm [9] is used to reduce the amount of training data. Codebooks are built for the both clients and non-clients of the system. The non-clients, i.e. the potential impostors, are represented by a reference model, which provides scores for counterbalancing the scores produced by the clients' models. Each target speaker is represented by codebook composed of 256-vectors.

2.4. The reference model design

A gender-specific reference model, which represents the set of potential impostor speakers, is created. As shown in Figure 1, a personal codebook for each background speaker, which is assumed a non-client, is built first. Each codebook consists of 256 vectors. These non-clients' codebooks are merged by gender, and separate reference models, collectively referred to as UBgCB are created for the male and female speakers. In the next step, the UBgCB size is reduced to 1024 vectors by using k-means clustering [9]. The UBgCB, along with the personal codebooks built for the enrolled speakers, are next employed to design an independent PNN for each target user.

2.5. The PNN-based classifier

The Probabilistic Neural Network (PNN) was chosen as a classifier for our speaker verification system, because of its good generalization properties and fast designing times. PNN are capable to learn from limited data, because their design is straightforward and does not depend on training [4]. As a result, PNNs are built only for a small fraction of the training time, which the back-propagation-trained neural networks require. Since the training of the PNNs is a non-iterative procedure, they are trained much faster than the state-of-the-art Gaussian Mixture Models (GMM)-based classifiers.

The PNNs, as introduced in [4] combine non-parametric probability density estimation with minimum risk decision making. The density estimation implements the Parzen window estimator by using a mixture of Gaussian basis functions. After the probability density functions for all classes are estimated, the posterior probabilities are computed, and then the Bayes' decision rule is applied to select the winning class. A comprehensive description of the PNN is offered in [4].

Since the speaker verification process is a two-class separation problem, a PNN for classification in $K = 2$ classes is considered. The probability density function $\mathrm{f}_i(\mathbf{x}_p)$ of each of the two classes κ_i, $i = 1,2$, is computed by:

$$\mathrm{f}_i(\mathbf{x}_p) = \frac{1}{(2\pi)^{d/2}\sigma^d}\frac{1}{M_i}\sum_{j=1}^{M_i}\exp\left(-\frac{1}{2\sigma^2}(\mathbf{x}_p-\mathbf{x}_{ij})^T(\mathbf{x}_p-\mathbf{x}_{ij})\right) \quad (1)$$

with $i = 1,2,$ where $\mathbf{x}_{ij}$ is the j- th training vector from class κ_i; $\mathbf{x}_p$ is the p- th input vector; d is the dimension of the speech feature vectors; and M_i is the number of training patterns in class κ_i. Each training vector $\mathbf{x}_{ij}$ is assumed a centre of a kernel function, and consequently the number of pattern units in the first hidden layer of the neural network is given as a sum of the pattern units for all classes. The standard devia-

tion σ acts as a smoothing factor, which softens the surface defined by the multiple Gaussian functions. As presented in equation (1), σ has the same value for all the pattern units. Therefore, a homoscedastic PNN is considered.

After the estimations of the class-conditional probability density functions is obtained through (1), the Bayesian decision rule (2) is applied to distinguish class κ_i, to which the p- th input vector $\mathbf{x}_p$ belongs:

$$\mathrm{D}(\mathbf{x}_p) = \underset{i}{\operatorname{argmax}} \left\{ h_i \; c_i \, \mathrm{f}_i(\mathbf{x}_p) \right\}, \; i = 1,2 \,, \tag{2}$$

where h_i is the a priori probability of occurrence of the patterns of category κ_i, and c_i is the cost function in case of misclassification of a vector belonging to class κ_i.

In general, the PNNs are easy to train, and retraining is easily carried out simply by adding new, or replacing existing pattern units in the first hidden layer of the structure. It is well known that the PNNs need more neurons compared to back propagation networks, which leads to an increased complexity and higher computational and memory requirements in the process of exploitation. On the other hand, the PNNs are based on parallel architecture, with independent processing units in the first and the second hidden layers, and therefore they are easy on implementation in hardware, when parallel computing devices are engaged. By exploiting this advantage, the PNN are capable to work much faster than the back propagation neural networks, which are not hardware friendly.

2.6. The score computation

The probability $P(k_i \mid \mathbf{X})$ all test vectors of a given test trial $\mathbf{X} = \{\mathbf{x}_p\}$, $p = 1,...,P$, to belong to class κ_i, is computed by:

$$P(k_i \mid \mathbf{X}) = \frac{N\left(\mathrm{D}(\mathbf{x}_p) = k_i\right)}{\sum_{j=1}^{K} N\left(\mathrm{D}(\mathbf{x}_p) = k_j\right)}, \quad i = 1,2, \tag{3}$$

where $N\left(\mathrm{D}(\mathbf{x}_p) = k_i\right)$ is the number of vectors $\mathbf{x}_p$ classified by the Bayesian decision rule (2) as belonging to class κ_i. Since the speaker verification task assumes an exhaustive taxonomy, any of the inputs $\mathbf{x}_p$ falls in one of the classes κ_i. Thus, the equality:

$$P = \sum_{j=1}^{K} N\left(\mathrm{D}(\mathbf{x}_p) = k_j\right), \tag{4}$$

where the number of test vectors P in the given trial $\mathbf{X}$, is always preserved.

For a given test trial, the averaged probability for all output decisions of a particular PNN, obtained by testing with multiple feature vectors, is utilized to compute a score:

$$\chi = \eta \cdot \left(P(k_1 \mid \mathbf{X}) - \beta \right), \tag{5}$$

where η and β are constants for tuning the scale and the offset of the produced score, respectively.

2.7. The final decision

A speaker-independent threshold is computed from a development set of data, in a manner that satisfies the decision strategy of the prospective application. Subsequently, the threshold θ is applied to the score (5), and a final decision $O(\theta)$ is made:

$$O(\theta) = \begin{cases} 1 & \text{for } \chi \geq \theta \\ 0 & \text{for } \chi < \theta \end{cases}. \tag{6}$$

When the score χ is above or equal to the threshold, the claimant user is accepted. Otherwise, the utterance is considered to belong to an impostor speaker.

2.8. The operational mode

In summary, the speaker verification system decides whether or not the trial belongs to the claimed speaker, depending on the degree of similarity of the input feature vectors to the speaker's model and to the reference model. Equation (1) estimates the similarity level by computing the corresponding Euclidean distances. For every input speech frame, a binary decision is made by applying the Bayesian decision rule (2). Next, through equation (3) an estimate of the probability a given test trial to belong to the claimed user is obtained. Finally, a speaker-independent threshold is applied to the score (5), and a final decision (6) is made

The PNN-based speaker verification system, described here, is implemented in MATLAB environment and the program code has not been optimized for speed. Nevertheless, it is capable of working in real-time on common personal computers, and uses about 35% of the resources of a Pentium 4 CPU working at 1.6 GHz.

3. The 2003 NIST SRE

In the 2003 NIST SRE, there were five main tasks: "One-speaker detection – cellular data", "Two-speaker detection – cellular data", "One-speaker detection – extended data", "One-speaker detection – multi-modal data", "Speaker segmentation – various data sources". These tasks were supported by individual datasets that included various training/testing conditions. Details are available in [1]. In the present work, we consider the "one-speaker detection – cellular data" task and the corresponding database.

The 2003 NIST SRE one-speaker detection database had been compiled by exploiting data from the huge Switchboard-Cellular Corpus, Part 2 and Part 3. Recordings from selected speakers, which had performed multiple sessions, calling over various public telephone networks in a number of environmental conditions, have been extracted to provide several representative sets. The phone calls had been performed form various devices and mismatching transmission channels. Each sub-set controls specific variables in such a way that analysis of the results from a single experiment can be analyzed form various perspectives. The speech files of the 2003 NIST SRE database had been processed accordingly to reduce any channel echoes and remove all significant speech pauses from the original recordings. A separation of caller-called channels also had been performed. Finally, the speech data for the one-speaker recognition task, disseminated to the participants in the evaluation, consisted of cellular speech of 146 male and 207 female speakers recorded in different environmental conditions, provisionally noted as: {'inside', 'outside', 'vehicle'}.

The training data consists of about two minutes of spontaneous speech, extracted from a single conversation. All training speech had been acquired over the mobile cellular networks of USA. The test trials are separated in five categories: {'00-15', '16-25', '26-35', '36-45', and '46-60'} seconds, depending on the amount of speech they contain. The primary condition task includes only these test trials, which contain

between fifteen and forty-five seconds of speech. The test data consist of speech recorded over {'TDMA', 'CDMA', 'Cellular', 'GSM', and 'Land'} transmission channels. The complete one-speaker detection task combines all tests. It included 2973 target and 34734 impostor trials. A detailed description of the 2003 NIST SRE is available in [1][11].

4. The 2003 NFI/TNO Forensic SRE

In brief, the 2003 NFI/TNO Forensic SRE targeted at assessing the practical value of the present state-of-the-art technology in a forensic context. Real-world wiretapped recordings, collected in real police investigations, from real criminal suspects, were used to provide conditions as close as possible to a real forensic procedure. This campaign aimed at evaluating speaker recognition systems from Universities, commercial companies, and research institutes. Some participants submitted results for multiple systems, but one (per institution) was identified as primary. One participating organization failed to provide results in the framework of the evaluation, and the results from another one did not meet the requirements. Results for the primary systems of the twelve participants that met the requirements were reported in [3][10].

The speech material of the 2003 NFI/TNO Forensic database [10] consists of real-field data, collected from recordings made using wiretaps for the purpose of police investigation. The recordings consist of wire-tapped cellular GSM to GSM telephone conversations recorded over a period of 23 months. All speakers are males. The telephone line quality varies between recordings from excellent to moderate (extremes at the lower end had been excluded). The telephone handsets used are unknown. The level and nature of background noises of the material varies and includes slight room reverberations, music in the background of the recording and in some cases background speakers (mostly children playing in the background). Although the speaking style is constant (spontaneous speech, laughter, shouting and whispering had been weed out) emotions varied between recordings from relaxed (frequent) to stressed (rare). The distribution of these parameters among speakers is not homogeneous. The range and distribution of recording dates between speakers varies. The material had been edited by NFI in order to select single speakers and to make the material anonymous. Signaling noises in the telephone recordings had been removed but speaking pauses had not been altered. The languages used are Dutch, English, Sranan Tonga (language spoken in Surinam) and Papiamento (language spoken at the Netherlands Antilles). A comprehensive account of the rules and procedure of the 2003 NFI/TNO Forensic Speaker Recognition Evaluation is available in [2].

In brief, the evaluation consisted of several separate experimental conditions, concentrating on different aspects of the speaker recognition problem. The main condition was a general performance evaluation, while other conditions investigated the influence of specific factors such as speech duration and spoken language. The training data available for building the target speaker models were 30, 60, and 120 seconds. The primary evaluation condition included the training data with duration of 60 seconds. The test trials had length of 7, 15, and 30 seconds. The complete evaluation consisted of 775 test fragments, each tested against eleven to sixty-six models out of a pool of 364 speaker models. In total, the test consisted of about 32000 trials. Details about the different tasks and evaluation rules are available in [2][10].

5. Discussion of the evaluation results

Summarizing the descriptions of the evaluation data provided in Sections 3 and 4 we can identify several differences between the 2003 NIST SRE and 2003 NFI/TNO Forensic SRE datasets and experimental conditions:

- Firstly, the 2003 NIST SRE deals with recordings collected from population of cooperative speakers, who are aware that their voice is recorded for research purposes. In opposite the database of the 2003 NFI/TNO Forensic SRE had been recorded via wiretapping and the speakers were not aware about the recording of their voice.
- The second difference is that the 2003 NIST SRE investigates mainly English language, while the 2003 NFI/TNO Forensic SRE deals with several languages and includes cross-language experiments.
- The third dissimilarity is that the 2003 NIST SRE doesn't name specific application, while the focus of the 2003 NFI/TNO Forensic SRE falls on resembling the conditions of real-world forensic procedure. By this reason, the 2003 NIST SRE offered a larger number of controlled experiments, which addressed a wider range of aspects of the real-world (for instance: gender, call location, transmission network, a wider range of trial durations).
- Moreover, the experimental conditions of the 2003 NIST SRE and 2003 NFI/TNO Forensic SRE were controlled in a dissimilar manner: in the 2003 NIST SRE the language and the amount of training data per target speaker were fixed, while in the 2003 NFI/TNO Forensic SRE these vary, but the gender of the speakers and the transmission standard were predetermined.

These distinctions, as well as the fact that the 2003 NIST SRE offered larger number of test trials in the separate experiments, made the 2003 NIST SRE and 2003 NFI/TNO Forensic SRE complementary to one another. Where parallels and resemblance of the experimental conditions between the two evaluations exist, they contribute for the better understanding of the specifics of the different setups.

Because the 2003 NFI/TNO Forensic SRE was carried out about six months after the 2003 NIST SRE, the WCL-1 system that participated in the both evaluations remained practically unchanged. This gives us the opportunity to investigate more details about the performance of our system and better understand its virtues and shortcomings.

Figure 2 and Table 1 present results for the WCL-1 system, which were obtained in the 2003 NFI/TNO Forensic SRE, and Figure 3 and Table 2 present these obtained in the 2003 NIST SRE. The figures present the DET plots for the various conditions, and the tables offer summary of the key numerical results and conditions for each specific subtask. Specifically, the Equal Error Rate (EER) in percentage and the actual decision cost (DCF_act) were complemented with the optimal decision cost (DCF_opt). While the EER gives intuitive, balanced, and application independent, but optimistic (due to the posteriori-computed threshold) assessment of the potential performance of a system, the DCF_act and DCF_opt are application-specific due to the *cost* coefficients C_{miss} and $C_{FalseAlarm}$ [1][2]. In particular, the DCF_opt gives impression about the prospective performance of a system if "the perfect" speaker-independent threshold is applied, while the DCF_act presents the actual performance of the system. In the 2003 NIST SRE and 2003 NFI/TNO Forensic SRE,

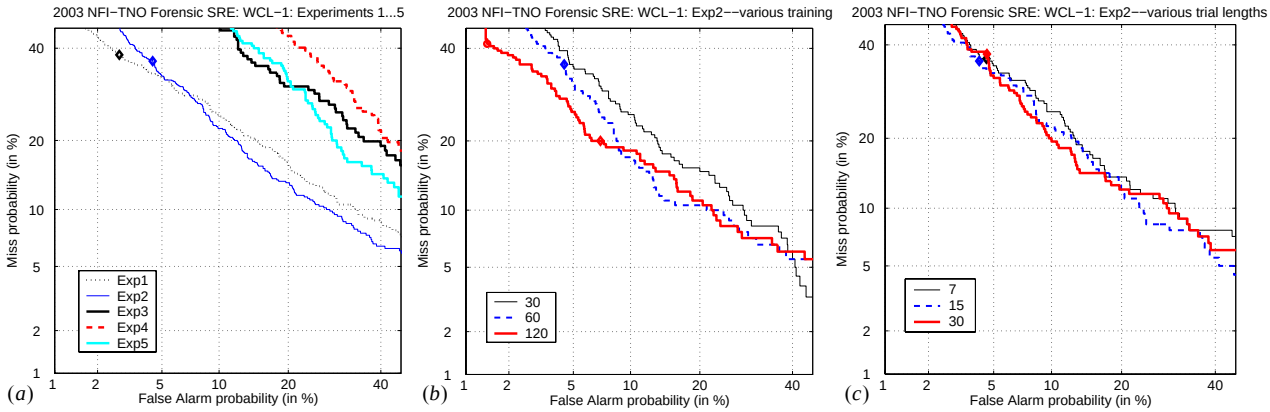

Figure 2. 2003 NFI-TNO Forensic SRE: Evaluation results for the WCL-1 system -- DET plots for: (*a*) Experiments 1 … 5, (*b*) experiment 2, depending on the amount of training data, (*c*) experiment 2, depending on the length of test trials.

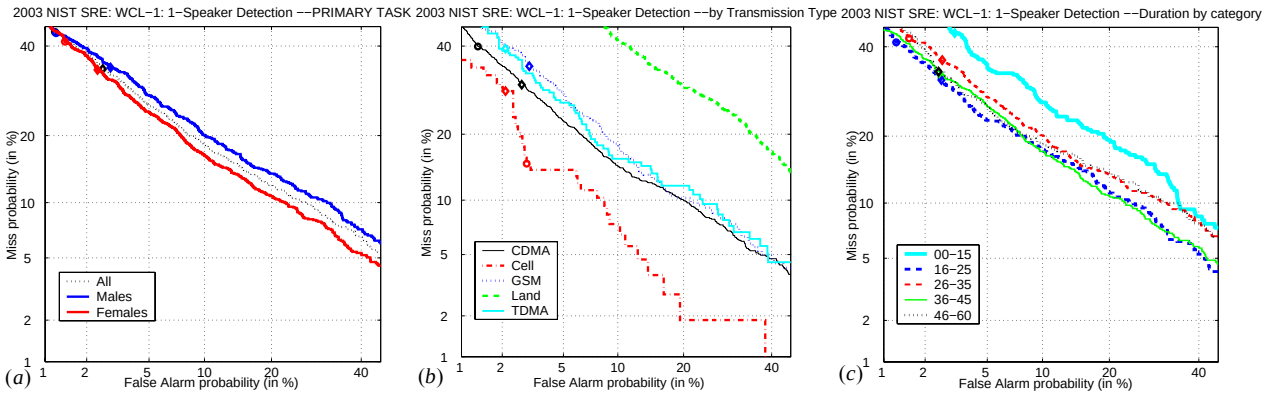

Figure 3. 2003 NIST SRE: Evaluation results for the WCL-1 system -- DET plots for: (*a*) Primary task, (*b*) depending on the transmission channel (primary trials), (*c*) depending on the length of test trials.

DCF_act was the formal performance measure according to which the participating systems were evaluated and ranked.

Figure 2(*a*) offers summary of experiments {1…5}, where Exp1 stands for the primary condition, Exp2 for the entire "Dutch" experiment (Dutch training and testing) with varying lengths of the training and test data. Exp3 has the settings of Exp1 except that the train language is English and the trials are non-English speech. Exp4 and Exp5 examine the effect of mismatch between train and test data for Dutch language: nonDutch models against Dutch trials, and Dutch models against nonDutch trials, respectively. Table 1 presents the numerical expression of the performance of the WCL-1 system for these subtasks, as well as for the Exp6, which evaluated a "proof in court" paradigm [10]. These result suggest, the WCL-1 system demonstrated its best performance in Exp2, mainly due to the availability of more training data (in some cases originating from different sessions) for a number of models – see Figure 2(*b*). Figure 2(*c*) illustrates the influence of test trial duration on the performance: the shortest and longest trials do not differ significantly. We deem this is because the WCL-1 system makes a binary decision (accept/reject) for each voiced speech frame. This lessens the influence of irrelevant speech frames. Next, these decisions are accumulated for the entire trial and after applying a speaker-independent threshold, a final decision is made. Thus, a trial with a longer duration brings a better precision of the confidence measure represented by (3), but this leads to a better final decision (6) only if the confidence is marginal to the speaker-independent threshold. The same holds for Figure 3(*c*), where the performance loss for the shortest duration tests (00-15 seconds) is mainly due to some trials for which voiced speech was not detected and by that reason, they were rejected. As Figure 2(*a*) presents, the mismatch between training and testing language (Exp3, Exp4, and Exp5) dramatically reduces the speaker verification performance when compared to the one of the primary condition, Exp1.

Figure 3(*a*) presents the performance of WCL-1 for the male and female speakers in the primary task of the 2003 NIST SRE. We deem the better performance for the female voices is mainly due to the larger amounts of voiced speech detected in the female training recordings. In average, the number of voiced speech frames in the female training data was about 70% larger that the number of voiced frames detected for the male speakers. In turn, this could due to the specifics of female voices, and our voiced/unvoiced detection.

Figure 3(*b*) illustrates the performance loss due to mismatch between training and testing for various transmission channels. As expected, the lowest EER and decision cost were observed for matching training and testing channels, i.e. testing data from type "Cell". Unfortunately, due to the small number of "Cell" trials, this experiment has a very low resolution, and precise performance estimation cannot be made.

Table 1. 2003 NFI-TNO Forensic Speaker Recognition Evaluation results for the WCL-1 system. The experimental conditions in the various tasks [10] are specified below the thin line.

	Exp1	Exp2	Exp3	Exp4	Exp5	Exp6
EER [%]	18.1%	15.7%	27.8%	31.5%	25.3%	20.2%
DCF act	**0.661**	0.806	1.000	1.000	1.000	0.639
DCF opt	0.613	0.682	0.832	0.919	0.939	0.694
# target trials	521	540	126	137	165	287
# impostor trials	9676	9360	2206	4648	2167	2353
size train data [s]	60s	30s,60s,120s	60s	60s	60s	60s
size test data [s]	15s	7s,15s,30s	15s	15s	15s	15s
train-test condition	Dutch-Dutch	Dutch-Dutch	English-nonEng	nonDutch-Dutch	Dutch-nonDutch	Dutch-Dutch

Table 2. 2003 NIST Speaker Recognition Evaluation results for the WCL-1 system. The experimental conditions in the various tasks [11] are specified below the thin line.

	Primary			All trials			Call location			Duration					Transmission				
	all	male	female	all	male	female	in	out	car	0-15	16-25	26-35	36-45	46-60	CDMA	Cell	GSM	Land	TDMA
EER [%]	14.8%	15.7%	13.8%	15.2%	15.8%	14.8%	16.2%	13.6%	12.0%	19.6%	14.8%	15.1%	14.3%	15.5%	12.8%	8.6%	13.4%	26.5%	14.5%
DCF act	**0.598**	0.632	0.575	0.607	0.637	0.588	0.627	0.548	0.572	0.747	0.574	0.625	0.582	0.582	0.574	0.498	0.651	0.855	0.602
DCF opt	0.553	0.552	0.550	0.568	0.561	0.570	0.584	0.511	0.539	0.692	0.534	0.574	0.538	0.576	0.531	0.440	0.568	0.774	0.580
# target trials	2215	924	1291	2973	1214	1760	2094	487	390	270	625	838	821	420	1271	107	361	468	178
# impostor trials	25944	10188	15756	34734	13517	21217	25053	5420	4241	3141	7328	9810	9628	4827	13975	1092	3995	7287	1901
size train data[s]	120s	120s	120s	120s	120s	120s	120s	120s	120s	120s	120s	120s	120s	120s	120s	120s	120s	120s	120s
size test data [s]	15-45s	15-45s	15-45s	0-60s	0-60s	0-60s	0-60s	0-60s	0-60s	0-15s	16-25s	26-35s	36-45s	46-60s	0-60s	0-60s	0-60s	0-60s	0-60s

The "CDMA", TDMA", and "GSM" trials led to similar results – a moderate decrease of performance was observed when compared to the no mismatch case. Finally, a dramatic increase of the EER and decision costs was observed for the trials originating from landline telephone network – roughly doubling of the EER, compared to the other cases. This is because in the WCL-1 system that participated in these two evaluations no special care was taken for normalization of the mismatch due to the non-linear transfer characteristics of handsets with carbon-button microphones and the landline transmission channel.

The results obtained for different call locations of the test trials (Table 2) show a lower performance for the "inside" location, when compared to "outside" and "car". We deem the prime cause for this effect are the reverberations from the walls of the room. However, other factors might contribute for the mismatch between test and train conditions as well.

Comparing the performance of WCL-1 between the primary tasks of the two SREs (see the column "Exp1" in Table 1, and "primary–male" in Table 2), we can recognize that "Exp1" was a more difficult task mainly due to the shorter length of the training speech. The same is valid for "Exp2" and the complete task "all trials–male" – a higher decision costs was obtained for "Exp2", mainly due to the smaller amount of training data for some of the target models.

6. Conclusion

The results obtained in the 2003 NIST SRE and 2003 NFI/TNO Forensic SRE pinpoint the shortcomings of the WCL-1 system evaluated here. Similarly to other systems, it is mostly suitable for applications where the user always makes the phone call from his/her personal mobile phone and communicates with the system on the language used during training of his/her client model. On another hand, among the main virtues of WCL-1 is the good competitiveness when compared to other systems (details in [3][10][11], where WCL-1 is system "15" in [10], and "T2" in [3] (a slightly modified version of WCL-1 is referred to as "S18" in [3])). It is of significant practical importance that the performance of WCL-1 does not depend heavily on the test trial duration.

7. Acknowledgement

We would like to thank to David van Leeuwen and Mark Przybocki for providing the Perl scripts, which divide site's result file in subtasks, and to Jos Bouten for the useful comments and corrections he made on an earlier version of our manuscript.

8. References

[1] "The NIST year 2003 speaker recognition evaluation plan". *National Institute of Standards and Technology of USA*, February 2003. Available: http://www.nist.gov/speech/tests/spk/2003/doc/2003-spkrec-evalplan-v2.2.pdf

[2] van Leeuwen, D., Bouten, J. "The NFI/TNO Forensic Speaker Recognition Evaluation Plan. Revision: 2.0," Sept. 2003. Available: http://speech.tm.tno.nl/aso/evalplan.pdf

[3] van Leeuwen, D., Martin, A.F., Przybocki, M.A., Bouten, J.S. "NIST and NFI/TNO evaluations of automatic speaker recognition", *Computer Speech and Language*, 2005. (*in press*)

[4] Specht, D.F. "Probabilistic Neural Networks". *Neural Networks,* 3(1):109–118, 1990.

[5] Slaney, M. "Auditory Toolbox. Version 2", *Technical Report #1998-010*, Interval Research Corporation, 1998.

[6] Ganchev, T., Fakotakis, N., Kokkinakis, G. "Comparative Evaluation of Various MFCC Implementations on the Speaker Verification Task", *Proc. of SPECOM 2005*, October 17-19, 2005. Patras, Greece, vol. 1, pp.191–194.

[7] Rabiner, L.R., Cheng, M.J., Rosenberg, A.E., McGonegal, C.A. "A comparative performance study of several pitch detection algorithms", *IEEE Trans. on Acoustics Speech and Signal Processing*, 24(5):399–418, 1976.

[8] Ganchev, T., Fakotakis, N., Kokkinakis, G., "Toward 2003 NIST Speaker Recognition Evaluation: The WCL-1 System", *Proc. of SPECOM'2003*, Moscow, Russia, October 27-29, 2003.

[9] Hartigan, J.A., Wong, M.A. "A k-means clustering algorithm", *Applied Statistics*, 28(1):100–108, 1979.

[10] van Leeuwen, D., Bouten, J. "Results of the 2003 NFI/TNO forensic speaker recognition evaluation". *Proc. Odyssey 2004 Speaker and Language recognition workshop*, *ISCA*, Toledo, Spain, 2004. pp. 75–82

[11] "The 2003 NIST Speaker Recognition Evaluation". Results CD, Workshop Presentations & Evaluation Results. *National Institute of Standards and Technology of USA*. Spring 2003 Evaluation. NIST Speech Disc r88_99_1