Intra-speaker variability effects on Speaker Verification performance

Juliette Kahn[1,2], *Nicolas Audibert*[1], *Solange Rossato*[2], *Jean-François Bonastre*[1]

[1]Laboratoire Informatique d'Avignon (LIA), University of Avignon, France
[2]Laboratoire Informatique de Grenoble (LIG), University of Grenoble, France
{juliette.kahn, nicolas.audibert, jean-francois.bonastre}@univ-avignon.fr, solange.rossato@imag.fr

Abstract

Speaker verification systems have shown significant progress and have reached a level of performance that make their use in practical applications possible. Nevertheless, large differences in terms of performance are observed, depending on the speaker or the speech excerpt used. This context emphasizes the importance of a deeper analysis of the system's performance over average error rate. In this paper, the effect of the training excerpt is investigated using ALIZE/SpkDet on two different corpora: NIST-SRE 08 (conversational speech) and BREF 120 (controlled read speech). The results show that the SVS performance are highly dependent on the voice samples used to train the speaker model: the overall Equal Error Rate (EER) ranges from 4.1% to 29.1% on NIST-SRE 08 and from 1.0% to 33.0% on BREF 120. The hypothesis that such performance differences are explained by phonetic contents of voice samples is studied on BREF 120.

1. Introduction

Over the last decade, automatic speaker verification systems (SVS) have been assessed regularly by the National Institute of Standards and Technology (NIST) [1]. The evaluation focuses on text-independent speaker detection and offers a common experimental protocol and a stable set of evaluation rules.

Although the task difficulty is changing through years, the NIST campaigns clearly show a drastic progress in terms of performance during the last years. The level of performance reached by the systems has become suitable for a large set of practical, commercial applications. Many applications are already available or planned for the next future, including the forensic ones. This context underlines the importance of a deep analysis of the system's performance, for instance on a per speaker basis, while the performance is usually assessed only through average error rate. In addition to the average performance information, performance variability also needs to be evaluated. Indeed, the identification of the performance variation factors is necessary for determining the contexts in which those systems may be used.

The system performance is commonly measured using two kinds of errors. A false acceptance (FA) occurs when an impostor is accepted by the system. A false rejection (FR) consists of rejecting a valid identity. Both error rates depend on the threshold used in the decision making process. Among the measures used to compare system performances, detection error trade-off (DET) curve [2], Equal Error Rate (EER) and Decision Cost Function (DCF) are usually used. The DET curve is obtained by plotting on a normal deviate curve the FA rate as a function of the FR rate. The EER corresponds to the operating point where FA rate = FR rate when the DCF corresponds to a specific operating point, described by the weight tied to each error (FA and FR) and the prior probabilities of these errors. In NIST evaluation, each training excerpt is regarded as produced by a different speaker, but the same speaker may have been recorded in several extracts. No comparison between these different extracts is conducted.

Some studies have investigated the possible causes of performance variation. [3] showed that the performance of the system may be improved by increasing the length of the training and testing signals. Indeed, the EER raised from 4.48% up to more than 30% when the duration of the training signals are shortened from 2.5 minutes to 10 seconds. Moreover, a short excerpt in training is more disadvantageous than a short excerpt in testing (more than 14% of EER with short excerpts in testing *vs.* more than 17% of EER with short excerpts in training)

Inter-speaker variation has also been studied. Doddington *et al.* [4] studied the errors induced by different speakers in 12 automatic speaker verification systems, and showed that the topology of the errors depends on speakers, consistently from one system to another. They distinguished 4 types of speakers, illustrated by a 'menagerie'. *Sheeps* correspond to the default speaker type (low FA, low FR). *Goats* are speakers who generate a disproportionate false rejection rate. *Lambs* correspond to speakers who generate a disproportionate false alarm rate. *Wolves* correspond to speakers that are likely to be mistaken for an other speaker.

Finally, the influence of the phonetic content of test excerpts was evaluated by [5]. Results suggest that glides and liquids together, vowels and more particularly nasal vowels and nasal consonants contain more speaker-specific information than phonetically balanced test utterances, even though the training excerpt were composed of 15 seconds of phonetically balanced speech.

This paper focuses on the variability due to the signal sample used to represent the speaker voice. The information about the speaker may differ among training excerpts. The aim of this paper is to quantify the effect of such a variability on SVS performance. The SVS scores for each training excerpt are compared in order to selected the best and the worst training excerpts. Global performance is assessed using two different databases.

Section 2 describes the system used. Section 3 and 4 investigate on the effect of training excerpt on NIST-08 and Bref 120 database respectively. A preliminary phonetic analysis on the BREF 120 database is conducted in section 5 before concluding in section 6.

2. System

The speaker verification system used in this paper is the open source toolkit ALIZE/SpkDet [6]. This system is regularly assessed during the NIST speaker recognition evaluation. It is based on the UBM/GMM approach and it includes a latent factor analysis inter-session variability modeling [7]. Since score normalizations show little effect on the performances, as illustrated by figure 1 (3.42%≤EER≤4.55%), no score normalization is applied .

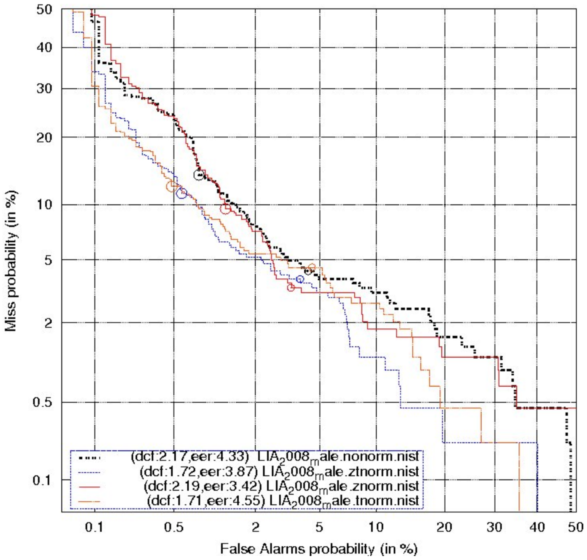

Figure 1: DET Curves without normalization and with ZT, Z and T normalizations (NIST08, male, english only).

3. Effect of the training excerpt on system performance on NIST-08

The aim of this part is to quantify the system performance range according to the training signal. A permutation training testing excerpt is conducted to evaluate the relative effects of training and testing on system's performances.

3.1. Experiments

3.1.1. *NIST-08*

The data collection used for experiments is derived from the male part of the *NIST-SRE08* telephone speech database. Most of the data are in English, but some conversations are collected in a number of other languages. The segment duration is approximately 2.5 minutes (condition short2-short3 in NIST protocol). This condition of the *NIST-SRE08* protocol is referred as *NIST-08* in this paper. A same speaker may have pronounced several training excerpt. *NIST-08* contains 221 speakers modeled from 648 training excerpts. 11,636 non-target trials and 874 target trials are conducted. Each speaker counts 4 target trials in average.

3.1.2. *M-08 : extension of NIST-08*

In order to maximize the number of target trials for each speaker, a leave-one-out scheme was implemented. For each speaker, a speaker model is trained using a speech sample while the other available samples of this speaker are used as target tests. This process is repeated for each speech segment available. This protocol is referred in this paper as *M-08*. 50 speakers of *NIST-08* pronounced less than two speech segments and are removed. *M-08* therefore includes 171 speakers with 816 excerpts, which means 3 to 15 voice samples per speaker. Each model computed from a given training excerpt was compared to 801 to 813 non-target tests, and to 2 to 14 target tests. As a result, a total of 661,416 non-target and 3,624 target trials are performed in *M-08*. Table 1 summarizes the number of speakers, models, target trials and non-target trials in *NIST-08* and *M-08* conditions.

	Speakers	Models	Target	Non-target
NIST-08	221	648	874	11,636
M-08	171	816	3,624	661,416

Table 1: Description of *NIST-08* and *M-08*.

3.1.3. *Best and worst models selection*

For each speaker, the best and the worst training files were selected among all the speech excerpts available for this speaker. For a given training file, FA and FR rates are estimated on *M-08*, using a threshold set to the EER point. This threshold is kept constant in all experiments performed on M-08. The best training excerpt is the one that minimizes FA+FR while the worst maximizes this value.
The average performances obtained with both training excerpts are compared to the average performance obtained using the training file defined in *NIST-08*. Speakers with only two speech excerpts were discarded from the analyzed set. In addition of that, the samples used as training excerpts for either the best or the worst speaker model were excluded from the test set, to avoid using a given file for both training and testing. Therefore, the test set was composed of the same speech signals for each training condition. These constraints give an experimental protocol with 511 target trials and 2,856 non-target trials.
In this protocol, 3 different conditions were applied in the selection of the training excerpt used to model each speaker :

- *NIST-3*. The training file is the one proposed in the original NIST protocol.
- *Min*. The training excerpt is selected by minimizing the sum of FA and FR rates computed on *M-08*.
- *Max*. The training excerpt is selected in order to maximize the sum of FA and FR rates computed on *M-08*.

3.1.4. *Training and test excerpts permutation*

Performance symmetry between testing signals and training signals is investigated to assess their relative weights in the performance obtained. If performance turns out to be symmetric, then errors may be explained by joint analyses of the training excerpt/test excerpt pairs. Conversely, large differences in performances obtained with the original pairs and the permuted ones would imply that training and tests excerpts have to be weighted when their characteristics are analyzed with regards to the performance induced.
NIST-03-inv, *Min-inv* et *Max-inv* are defined as the symmetric sets of *NIST-03*, *Min* and *Max* respectively. In these 3 permuted sets, training excerpts of the original sets are used for testing, and testing excerpts for training.

3.2. Results

3.2.1. *Training excerpt effect*

Figure 2 presents the DET curves for the 3 conditions (NIST-3, Min, Max). The EER are 12.1%, 4.1%, and 21.9% for NIST-3, Min, and Max conditions respectively. Looking at these results, it appears clearly that the choice of the training excerpt used to model each speaker plays an important role in the performance of the speaker verification system.

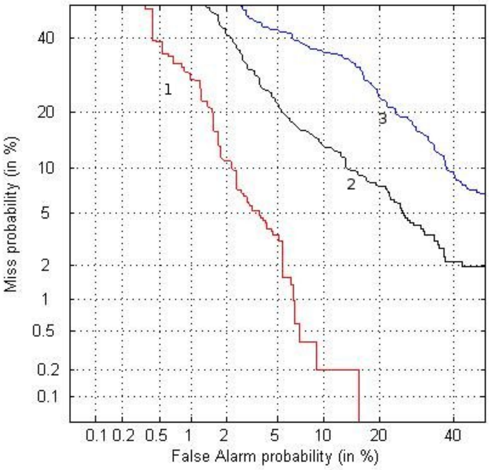

Figure 2: DET curves for *Min* (EER=4.1%) (1), *NIST-3* (EER=12.1%) (2), and *Max* (EER=21.9%) (3).

3.2.2. *Permutation*

Figure 3 presents DET curves for each set *NIST-3-inv*, *Min-inv* and *Max-inv*. When compared to corresponding non-inverted sets, the EER *Min-inv* raises up to 7,4% (+ 3.1%), while the EER in *Max-inv* decreases down to 17,0% (- 4.9%). The EER for *NIST-3* is 13,5% (+ 1.4%). The difference between the original set and the permuted one is substantial (more than 3 points) in the case of worst and best models.

3.3. Discussion

Training signal selection substantially modifies the global performance of the system while the speakers and the test excerpts remain the same in each set. Indeed, while the performance in the *Min* set is better than in the *Max* set, this pattern is reverted when training and test signals are permuted to obtain the *Min-inv* and *Max-inv* conditions. It is worth noting that, even if the ranking is the same, the variation in performance is higher when the training excerpts vary than when the testing excerpt vary.

The number of frames selected in *Min* and *Max* excerpts are significantly different, as shown by a paired t-test (t(170)=11.11, p<0.001). The training excerpts that lead to the best speaker verification performance contain more frames (7495 in average in *Min vs.* 7112 in *Max*). However, other factors may account for the performance differences observed between those two sets of models. Indeed, the NIST database includes different langugages and different recording conditions. In addition of that, the phonetic content may also vary among files.

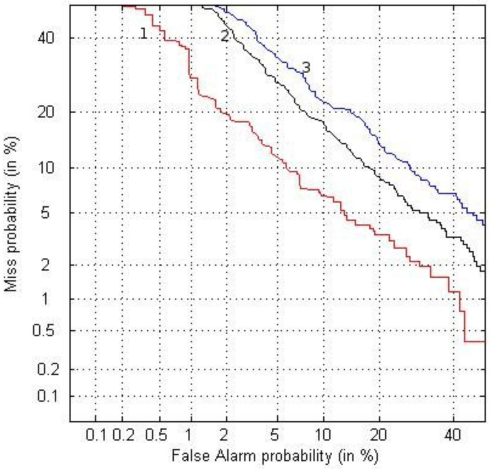

Figure 3: DET curves for *Min-inv* (EER=7.4%) (1), *NIST-3-inv* (EER=13.5%) (2), and *Max-inv* (3) (EER=17.0%).

4. Effect of the training excerpt on system performance on BREF 120

The BREF 120 database contains 66,000 single-session phonetically balanced aloud-read French sentences [8]. The available transcriptions may be used in order to obtain a phonetic labeling.

4.1. Experiments

4.1.1. *BREF 120*

The BREF 120 database is mainly composed of sentences produced by native French speakers, but also includes non-native speakers that have been discarded from the present study. The 64 female and 47 male remaining French native speakers were considered in this experiment. For each speaker, sentences were concatenated in a random way in order to generate files that contain a number of selected frames bigger than 3000, i.e. more than 30 seconds of selected speech signal. As an integer number of sentences are concatenated without being cut, the number of selected frames varies from 3400 up to 4200 per file. A set of 39 files is generated for each speaker. 18 files are reserved for training while the 21 files are used for testing. All combinations are assessed. For each male speaker, 378 target trials and 17,388 non-target trials are conducted. For each female speaker, 378 target trials and 23,814 non-target trials are conducted. Altogether, more than 2,383,000 trials are conducted

For the sake of comparability with *NIST-08*, longer files of 2.5 minutes are also used. In this condition, only 3 files are used for training and 3 other files for testing for each speaker. There are 576 and 423 target trials and 36,288 and 19,458 non-target trials for female and male speakers respectively. Table 2 summarizes the numbers of trials for 2.5 minutes-long and 30 seconds-long files.

	Speakers	Models	Target	Non-target
F 2.5 minutes	64	192	576	36,288
M 2.5 minutes	47	141	423	19,458
F 30 seconds	64	1,152	24,192	1,524,096
M 30 seconds	47	846	17,776	817,236

Table 2: Description of BREF 120 for 2.5 minutes-long and 30 seconds-long files. F: female speakers; M: male speakers.

4.1.2. *Best and worst model selection*

For each set (30 seconds and 2.5 minutes), the best and the worst training files of each speaker were selected among the available speech files for training. For a given training file, FA and FR rates were estimated on all the testing files for 30 seconds on the one hand and 2.5 minutes on the other hand. The best excerpt minimizes FA+FR while the worst one maximizes this value. The best excerpt for each male speaker was stored in the set of training files *Min-males*, and the worst one in *Max-males*. Similarly, best and worst excerpt for each female speaker were stored in *Min-females* and *Max-females*. As a result, *Min-males* and *Max-males* count 47 files, while *Min-females* and *Max-females* count 64 files. The performance obtained for those 4 sets of files is compared to the performance of random sets *Random-males* and *Random-females*, obtained by randomly selecting 10 training files per speaker. All the trials are performed with the same set of testing files.

4.1.3. *ALIZE/SPkDet parametrization*

French data from ESTER [9] were used to parameterize the male and female UBM. Considering the recording condition (single session, aloud reading), no factor analysis was performed.

4.2. Performance of ALIZE/SPkDet on BREF 120

4.2.1. *Global performance*

Figure 4 presents the DET curves obtained from all available training files, separately for male and female speakers. The performance variations observed are bigger in the *30 seconds* condition than in the *2.5 minutes* condition as shown in table 3. For 2.5 minutes-long files, EER are 2.3% and 2.8% for female and male speakers respectively. When the files are shorter, the EER raise up to 9.9% and 8.8% respectively. These rates are similar to those obtained in others studies for similar lengths [3]. For both *30 seconds* and *2.5 minutes* conditions, EER remain lower than those obtained on 2.5 minutes-long files with the NIST-08 database. However, it should be noted that the training files used in the NIST-08 database cannot be directly compared to the *2.5 minutes* condition defined on the BREF database, since the latter include 2.5 minutes of selected frames *vs.* 2.5 minutes of speech (2.0 minutes of selected frames in average on *Min* and *Max* sets) in NIST-08.

4.2.2. *Worst and best models*

Considering only the best training files in the *2.5 minutes* condition, the EER decreases to 0.4% and 0.9% for female speakers (*Min-female*) and male speakers (*Min-male*) respectively. Considering only the best training files, it increases to 5.3% for *Max-female* and *Max-male* sets. Figures 7 and 8 presents the DET curves for *Max* and *Min* in the *2.5 minutes* condition, respectively for male and female speakers.

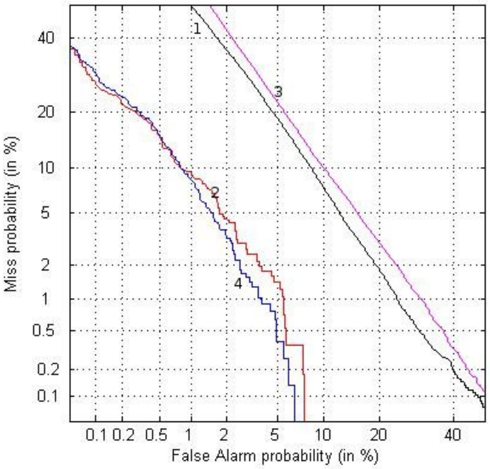

Figure 4: DET Curves for *Male-30 seconds* (EER=8.8%) (1), *Male-2.5 minutes* (EER=2.8%) (2), *Female-30 seconds* (EER=9.9%) (3), *Female-2.5 minutes* (EER=2.3%) (4).

In *30 seconds* condition, while the EER is about 9% when all the files are taken into account, this value decreases to 1% when the best training files are selected. On the contrary, when the worst models are selected. EER increases up to 28.5% and 33.0% for *Max-females* and *Max-males* respectively. The EER of the 10 random selections ranges from 8,8% to 11,5% (mean=10.3%, standard deviation=1.1) for female speakers and from 6.3% to 11.6% (mean=9.0%, standard deviation=1.4) for male speakers. These mean values of EER are close to the global EER obtained for all the training files. As shown on figures 5 and 6, the variation of performances among random selections remains far beyond the difference observed between *Min* and *Max* conditions. Similarly to the results obtained on the NIST-08 database, large performance variations due to the choice of the training file are observed on the BREF database. Indeed, substantial EER differences between *Min* and *Max* sets are found both in the *2.5 minutes* (about 30% difference) and *30 seconds* (about 4.5% difference) conditions.

	Global	Min	Max	Random
F 2.5 minutes	2.3%	0.4%	5.3%	-
M 2.5 minutes	2.8%	0.9%	5.3%	-
F 30 seconds	9.9%	1.1%	28.5%	10.3% (1.1%)
M 30 seconds	8.8%	1.0%	33.0%	9.0% (1.4%)

Table 3: Summary of EER obtained on the BREF 120 database, for the *Min*, *Max* and *Random* sets in *2.5 minutes* and *30 seconds* conditions. Values in parentheses for the *Random* set indicate the EER standard deviation on the 10 selected training files. F: female speakers; M: male speakers.

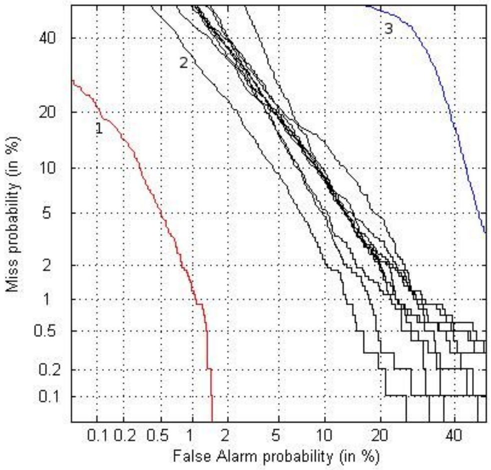

Figure 5: DET curves for *Min-male* (EER=1.0%) (1), 10 random sets for male speakers (6.3%<EER_Random<11.6%) (2), and *Max-male* (EER=3.0%) (3) in the *30 seconds* condition.

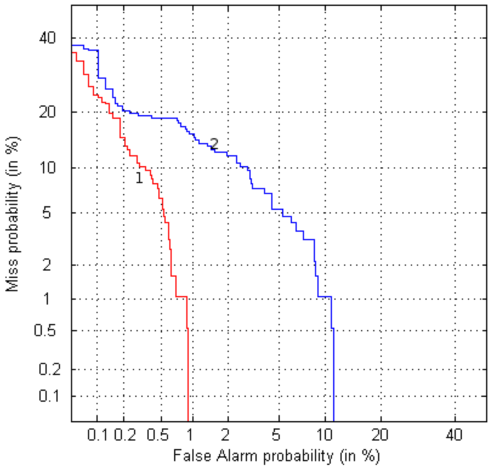

Figure 7: DET curves for *Min-male* (EER=0.9%) (1) and *Max-male* (EER=5.3%) (2) in the *2.5 minutes* condition.

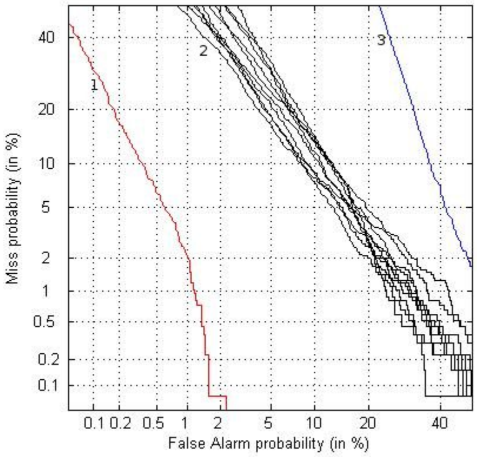

Figure 6: DET curves for *Min-female* (EER=1.0%) (1), 10 random sets for female speakers (6.3%<EER_Random<11.6%) (2), and *Max-female* (EER=3.0%) (3) in the *30 seconds* condition.

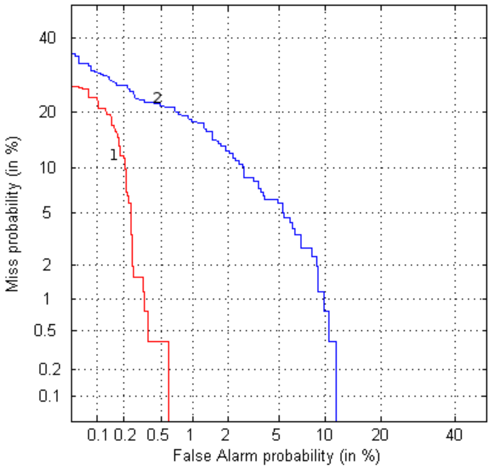

Figure 8: DET curves for *Min-female* (EER=0.4%) (1) and *Max-female* (EER=5.3%) (2) in the *2.5 minutes* condition.

5. Preliminary phonetic analysis

A possible factor for speaker verification performance variation may be the phonetic content of the excerpts. Indeed, it has been shown that some phonetic segments were more efficient than others for modeling a speaker [5]. This section presents some preliminary analysis of phonetic content of *Min* and *Max* sets. Since system performances were evaluated separately for male and female speakers, a distinct phonetic analysis is carried out for each speaker gender.

5.1. Phonetic transcription

A forced alignment of the speech signals is obtained using the open-source toolbox Speeral [10]. A manual modification of the phonetized lexicon was necessary to obtain a correct alignment of the BREF sentences. To ensure a full coverage of the words produced in BREF, entries corresponding to abbreviations, acronyms, proper names, etc., were added to this lexicon together with allophonic variants. The phonetic alignment was manually checked on a randomly selected subset. Only the frames selected by the speaker verification system were taken into account. An additional class was necessary for the frames selected during non speech part (NS). The forced-alignment software of the Speeral toolbox has a single acoustic model for each of the pairs of vowels /a-ɑ/, /ø-ə/ and /ɛ̃-œ̃/, which opposition is neutralized in standard French, except in regional varieties not represented in the BREF database. Moreover, it considers nasal consonants /ɲ/ and /ŋ/ as realized respectively as [nj] and [nɡ]. As a result, the phonetic labeling obtained associates each frame with one of the 32 remaining phonemes of the French language or to the NS class.

5.2. Phonetic distribution

5.2.1. Descriptive statistics

The phonetic distributions of *Max*, *Min* and *Random* are compared in order to evaluate the kind of information provided to the speaker verification system. The number of selected frames is calculated for each phoneme and NS. These distributions are extracted from the *30 seconds* files. The first analysis concerns the possible difference between *Min* file distributions and *Max* file distributions. The mean distribution of the 10 *Random* files for each speaker is also analyzed. For comparison, the 21 test files of each speaker are also analyzed and the mean distribution is computed. Figure 9 presents the phoneme distributions for *Max*, *Min* and *Random* and *Test* sets for female and male speakers.

Distributions show similar patterns. Indeed, the ten more frequent phonemes are /s ʁ a i e ɛ l t ɑ̃ d/ which almost matches French phoneme frequencies given by [11]. Large variations are observed among speakers, as illustrated by standard deviation values. Small differences are observed between the distributions of *Max*, *Min* and *Random* and *Test*.

5.2.2. Statistical comparison of Min, Max and Random sets

Differences between *Max*, *Min* and *Random* are evaluated by means of statistical hypothesis tests. Hypothesis tests used in the present study evaluate the probability p that observed differences between groups defined by a categorical factor occurred by chance, using Fischer's F-distribution (see for instance [12] for details about statistical hypothesis tests and their application). The difference is considered as significant if the obtained p-value is below an arbitrary threshold, classically set to 0.05 in behavioral and speech sciences. The most widely used hypothesis tests are Student's t-test and univariate or multivariate analyses of variance (ANOVA). While univariate analyses compare observations described by a single numeric dependent variable, multivariate analyses compare observations described by a set of numeric dependent variables. Repeated-measures designs make possible such comparisons while taking into account the inter-subjects variability, by grouping measures obtained from the same subject at the same level.

The statistical difference between phonetic distributions in *Max*, *Min* and *Random* was evaluated using a repeated-measures multivariate analysis of variance (repeated-measures MANOVA, see for instance [13]) for each speaker gender. The numbers of frames of each phoneme were defined as dependent variables, and the training file group (*Max*, *Min* or *Random*) as between-subject factor. Results indicate that the phonetic distribution of *Max*, *Min* and *Random* do not significantly differ, neither for female (F(33,31)=1.40, p=0.175) nor for male speakers (F(33,14)=2.23, p=0.056).

Differences of total number of selected frames in *Max*, *Min* and *Random* were evaluated by an univariate ANOVA with the number of selected frames as dependent variable and the performance group as independent factor. Contrary to the results obtained with the *NIST-08* database, the total number of selected frames does not significantly differ from one performance group to another (female speakers: F(2,189)=2.34, p=0.099; male speakers: F(2,138)=1.12, p=0.331).

Comparison of the number of frames of each phoneme in *Max vs. Min* models was performed using an univariate repeated-measures ANOVAs per phoneme, separately for male and female speakers. Phonetic distributions in *Max vs. Min* for female speakers significantly differ only on the quantity of /t/, bigger in *Min* (F(1,63)=4.84, p=0.032). For male speakers, those distributions differ on the quantity of /k/, bigger in *Max* (F(1,46)=7.58, p=0.008) and on the quantity of /ʁ/, bigger in *Min* (F(1,46)=5.83, p=0.020).

Repeated-measures MANOVAs comparing distribution in phonetic classes of *Max*, *Min* and *Random* were also performed. Phonetic classes were defined as indicated on figure 9. Those analyses yield the same global results as the comparison of distributions in phonemes (female speakers: F(10,54)=1.92, p=0.062; male speakers: F(10,37)=1.59, p=0.149). Comparison of the number of frames of each phonetic class in *Max vs. Min* indicates that only nasal consonants in female speakers' models are significantly more represented in *Max* (F(1,63)=4.38, p=0.040), while other phonetic classes are equally represented. Such limited differences in phonetic distributions between *Max vs. Min* can hardly account for the large differences in system performances observed between those conditions. Those results therefore suggest that system performance variability might be better explained by differences in intrinsic acoustic quality of speech segments.

5.3. Analysis of acoustic features

The coefficients used by the speaker verification system are the normalized LFCC, Delta and Delta-Delta. In this analysis we consider LFCC, Delta, Delta-Delta separately. LFCC values provide information on the spectral characteristics of phonemes while Delta and Delta-Delta values reflect dynamic information. For each phoneme, coefficients extracted from all frames in *Min* and *Max* were compared using a MANOVA, separately for male and females speakers. The 20 coefficients corresponding to LFCC, Delta or Delta-Delta were set as dependent variables, and the training file group (*Max* or *Min*) as independent factor. Table 4 summarizes the statistical significance of the comparison of *Max vs. Min* obtained in the 132 MANOVAs performed (33 phonemes x 3 sets of 20 coefficients x 2 speaker genders).

Except the /v/ produced by female speakers, LFCC significantly differ in *Max* vs. *Min* for all phonemes. Delta values significantly differ for 33% of phonemes, with limited match between female and male speakers except for nasal consonants and most plosives. This result is in line with studies by [14], who pointed out formant transitions, supposedly described by Delta coefficients, as carrying information on the speaker. Values of Delta-Delta coefficients are not significantly different in *Max* vs. *Min*, except for /j/ produced by both female and male speakers, as well as /l/ of female speakers, and /a/ and /i/ of male speakers.

6. Conclusions

Experiments were carried out in order to determine the role of the training excerpt. The EER sensitivity to the training material was quantified by evaluating EER on specifically built subsets of the two corpora analyzed. This quantization was achieved by selecting the training excerpts of each speaker that produce the largest and smallest EER, compared to a baseline condition obtained by randomly selected the excerpts used to model each speaker. A large EER variation was observed depending on the choice of the training excerpt used to model each speaker. Indeed, the EER range from 4.1% to 21.9% according to the voice sample selected for the speaker model

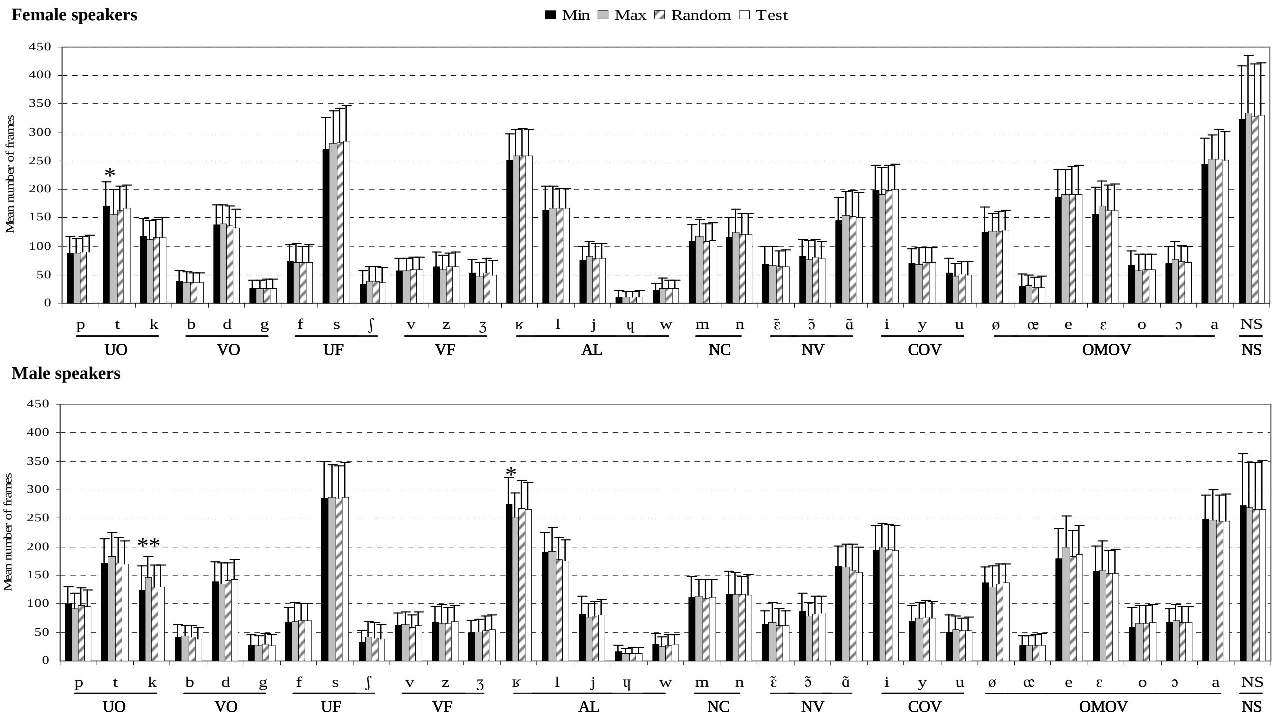

Figure 9: Phonetic distribution in *Min*, *Max*, *Random* and *Test* sets for female (top) and male (bottom) speakers. Error bars represent standard deviation. Stars indicate phonemes for which the number of frames is significantly different in *Min vs. Max*: **: p<.01; *: p<.05. Phonetic classes are indicated below phonemes labels. UO, VO: unvoiced and voiced occlusive consonants. UF, VF: unvoiced and voiced fricative consonants. AL: approximant and lateral consonants. NC: nasal consonants. NV: nasal vowels. CV: closed vowels. OMV: opened and median vowels.

in the NIST-SRE08 database, and from about 1% to 30% in BREF 120 with 30 seconds-long excerpts. For a given speaker, the characteristics of the training sample therefore turns out to be an important factor that may explain the performance variations of a speaker verification system. Moreover, the large variation observed when permuting training and test signals indicates that those sets of signals have to be differentiated in the analysis.

The question of the relevant phonetic information for speaker verification was addressed, and preliminary analyses were performed on the BREF 120 corpus. Despite a large variation among speakers, no significant difference was found between the phoneme distributions of the training files that generate few errors and those that generate a lot of errors. Observed phonetic distributions seem to be more representative of the French phoneme frequencies than of specific training signals. Since the utterances of the BREF database were chosen as phonetically balanced, this result is not surprising. Although it suggests that differences in phonetic distributions cannot account alone for all the performance variability observed, it does not imply that the phonetic distribution of the training excerpt has no effect on the speaker verification performance.

In order to further investigate properties of training signals that may explain performance differences, the acoustic qualities of phonemes whithin a given phonetic label were compared in the best *vs.* worst training files in terms of EER. As a first step, cepstral coefficients used by the system were compared. Significant differences were found between the two sets of training files for LFCC coefficients (supposed to describe the segmental information) extracted from the majority of phonemes. Delta coefficients supposed to carry information on transitions also differ significantly in the two sets for part of the phonemes represented, especially for occlusive consonants, including nasals. Conversely, acceleration information of Delta-Delta coefficients very seldom differ between best and worst training files.

Further work is needed to determine what kind of acoustic information induces high speaker verification performances, and what kind of information damages SVS performances. Indeed, for voiced segments, within-phoneme differences in cepstral information may reflect either variation in supralaryngeal settings or in voice quality. Since formant transitions have been shown to carry information on the speaker [14], the analysis of the link between this information and cepstral coefficients is of particular interest.

Determining the link between acoustic information and speaker verification performances would make possible the development of a confidence measure on models. Such a confidence measure would enable the prediction of the SVS performance on a given model.

Finally, a per speaker analysis may be also conducted. Indeed, although the phonetic distributions in best *vs.* worst training files do not significantly differ, a large inter-speaker variability was observed. The dispersion of the phonetic inventory of each speaker may therefore bring interesting information, as well as the inter-speaker variability in acoustic properties of excerpts. Such studies are needed to determine phonetico-acoustic profiles of the different animals that compose Doddington's menagerie.

	F			M		
Phon.	LFCC	Delta	D-D	LFCC	Delta	D-D
p	*****	****	n.s.	**	**	n.s.
t	******	**	n.s.	******	****	n.s.
k	****	n.s.	n.s.	****	***	n.s.
b	******	****	n.s.	******	*	n.s.
d	***	**	n.s.	***	*****	n.s.
ɡ	******	*	n.s.	******	***	n.s.
f	******	**	n.s.	***	n.s.	n.s.
s	****	***	n.s.	*****	**	n.s.
ʃ	******	n.s.	n.s.	******	n.s.	n.s.
v	n.s.	**	n.s.	***	**	n.s.
z	***	n.s.	n.s.	******	*	n.s.
ʒ	******	n.s.	n.s.	******	n.s.	n.s.
ʁ	***	**	n.s.	******	***	n.s.
l	***	n.s.	*	******	***	n.s.
j	******	*	*	******	n.s.	*
ɥ	******	*	n.s.	******	n.s.	n.s.
w	******	**	n.s.	****	n.s.	n.s.
m	******	****	n.s.	******	*	n.s.
n	******	******	n.s.	******	*	n.s.
ɛ̃	******	*	n.s.	******	n.s.	n.s.
ɔ̃	*	**	n.s.	******	**	n.s.
ɑ̃	******	n.s.	n.s.	******	***	*
i	*****	**	n.s.	******	*****	**
y	****	n.s.	n.s.	******	n.s.	n.s.
u	******	n.s.	n.s.	******	*	n.s.
ø	******	*	n.s.	******	n.s.	n.s.
œ	******	****	n.s.	*	n.s.	n.s.
e	******	***	n.s.	******	**	n.s.
ɛ	******	n.s.	n.s.	******	****	n.s.
o	******	**	n.s.	******	**	n.s.
ɔ	******	n.s.	n.s.	***	******	n.s.
a	******	*****	n.s.	**	**	n.s.
NS	***	n.s.	n.s.	***	n.s.	n.s.

Table 4: Significance of the comparison of MFCC, Delta and Delta-Delta (D-D) coefficients in *Max vs. Min* for each speaker gender and each phoneme. ******: p<.000001; *****: p<.00001; ****: p<.0001; ***: p<.001; **: p<.01; *: p<.05; n.s.: non significant.

Acknowledgment. This study was partly funded by the French *Direction Générale de l'Armement*, and by the European project MoBio[1] (FP7-2007-ICT-1).

7. References

[1] M. Przybocki and A. Martin, "NIST speaker recognition evaluation chronicles.," in *Odyssey 2004 Workshop*, Toledo, Spain, June 2004.

[2] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. A. Przybocki, "The det curve in assessment of detection task performance," in *Eurospeech '97*, Rhodes, Greece, Sept. 1997.

[3] B. Fauve, N. Evans, and J. Mason, "Improving the performance of text-independent short duration SVM-and GMM-based speaker verification," in *Odyssey 2008 Workshop*, Stellenbosch, South Africa, Jan. 2008.

[4] G. Doddington, W. Liggett, A. Martin, M. Przybocki, and D. Reynolds, "Sheep, goats, lambs and wolves: A statistical analysis of speaker performance in the NIST 1998 speaker recognition evaluation.," in *ICSLP 98*, Sydney, Australia, Dec. 1998.

[5] I. Magrin-Chagnolleau, J.-F. Bonastre, and F. Bimbot, "Effect of utterance duration and phonetic content on speaker identification using second-order statistical methods," in *Eurospeech '95*, Madrid, Spain, Sept. 1995.

[6] J.-F. Bonastre, N. Scheffer, D. Matrouf, C. Fredouille, A. Larcher, A. Preti, G. Pouchoulin, and N. Evans, "ALIZE/SpkDet : a state-of-the-art open source software for speaker recognition," in *Interspeech 2007*, Antwerp, Belgium, Aug. 2007.

[7] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Factor analysis simplified," in *ICASSP '05*, Philadelphia, PA, USA, Mar. 2005, vol. 1, pp. 637–640.

[8] L.F. Lamel, J.-L. Gauvain, and M. EskEnazi, "BREF, a large vocabulary spoken corpus for French," in *Eurospeech'91*, Genova, Italy, Sept. 1991.

[9] G. Gravier, J. F Bonastre, S. Galliano, E. Geoffrois, K. Mc Tait, and K. Choukri, "The ESTER evaluation campaign of rich transcription of french broadcast news," in *4^{th} LREC*, Lisbon, Portugal, May 2004.

[10] G. Linares, P. Nocera, D Massonie, and D. Matrouf, "The lia speech recognition system: from 10xrt to 1xrt," *Lecture Notes in Computer Science*, vol. 4629, pp. 302–308, 2007.

[11] B. Tranel, *The sounds of French: An introduction*, Cambridge University Press, 1987.

[12] T. Rietveld and R. van Hout, *Statistics in language research: analysis of variance*, Mouton de Gruyter, 2005.

[13] K. P Weinfurt, "Repeated measures analyses: ANOVA, MANOVA, and HLM," in *Reading and understanding MORE multivariate statistics*, L. G. Grimm and P. R. Yarnold, Eds. pp. 317–361, American Psychological Association, 2000.

[14] K. McDougall and F. Nolan, "Discrimination of speakers using the formant dynamics of /u:/ in British English," in *16^{th} ICPhS*, Saarbrücken, Germany, Aug. 2007.

[1] http://www.mobioproject.org/