Signal Processing 88 (2008) 2956–2970

Contents lists available at ScienceDirect

Signal Processing

journal homepage: www.elsevier.com/locate/sigpro

Fast and accurate sequential floating forward feature selection with the Bayes classifier applied to speech emotion recognition

Dimitrios Ververidis, Constantine Kotropoulos *

Artificial Intelligence and Information Analysis Laboratory, Department of Informatics, Aristotle University of Thessaloniki, Box 451, Thessaloniki 541 24, Greece

ARTICLE INFO

Article history:
Received 11 May 2007
Received in revised form
19 June 2008
Accepted 3 July 2008
Available online 5 July 2008

Keywords:
Bayes classifier
Cross-validation
Variance of the correct classification rate of the Bayes classifier
Feature selection
Wrappers

ABSTRACT

This paper addresses subset feature selection performed by the sequential floating forward selection (SFFS). The criterion employed in SFFS is the correct classification rate of the Bayes classifier assuming that the features obey the multivariate Gaussian distribution. A theoretical analysis that models the number of correctly classified utterances as a hypergeometric random variable enables the derivation of an accurate estimate of the variance of the correct classification rate during cross-validation. By employing such variance estimate, we propose a fast SFFS variant. Experimental findings on Danish emotional speech (DES) and speech under simulated and actual stress (SUSAS) databases demonstrate that SFFS computational time is reduced by 50% and the correct classification rate for classifying speech into emotional states for the selected subset of features varies less than the correct classification rate found by the standard SFFS. Although the proposed SFFS variant is tested in the framework of speech emotion recognition, the theoretical results are valid for any classifier in the context of any wrapper algorithm.

1. Introduction

Vocal emotions constitute an important constituent of multimodal human computer interaction [1,2]. Several recent surveys are devoted to the analysis and synthesis of speech emotions from the point of view of pattern recognition and machine learning as well as psychology [3–6]. The main problem in speech emotion recognition is how reliable is the correct classification rate (CCR) achieved by a classifier. This paper derives a number of propositions that govern the estimation of accurate CCRs, a topic that has not been addressed adequately, yet.

The classification of utterances into emotional states is usually accomplished by a classifier that exploits the acoustic features that are extracted from the utterances. Such a scenario is depicted in Fig. 1. Feature extraction consists of two steps, namely the extraction of acoustic feature contours and the estimation of global statistics of feature contours. The global statistics are useful in speech emotion recognition, because they are less sensitive to linguistic information. These global statistics will be called simply as features throughout the paper. One might extract tens to thousands of such features from an utterance. However, the performance of any classifier is not optimized, when all features are used.

Indeed, in such a case, the CCR, usually deteriorates. This problem is often known as 'curse of dimensionality', which is due to the fact that a limited set of utterances does not offer sufficient information to train a classifier with many parameters weighing the features. Therefore, the use of an algorithm that selects a subset of features is necessary. An algorithm that selects a subset of features, which optimizes the CCR, is called a *wrapper* [7].

Different feature selection strategies for wrappers have been proposed, namely *exhaustive*, *sequential*, and *random search* [8,9]. In exhaustive search, all possible combinations of features are evaluated. However, this method is

* Corresponding author. Tel./fax: +30 2310 998225.
E-mail addresses: jimver@aiia.csd.auth.gr (D. Ververidis), costas@aiia.csd.auth.gr (C. Kotropoulos).

0165-1684/$ - see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.sigpro.2008.07.001

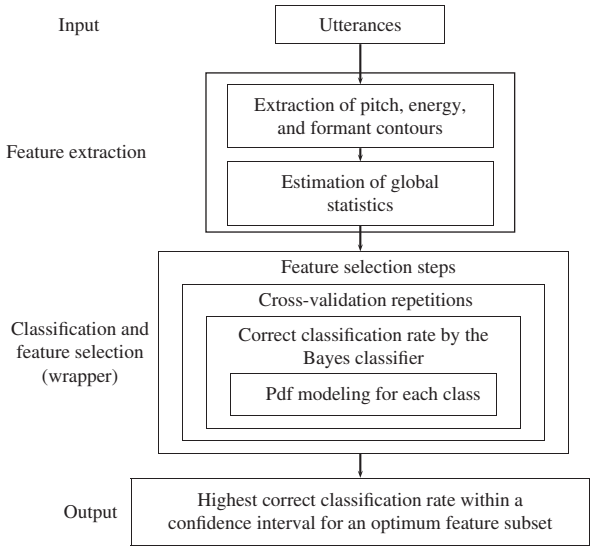

Fig. 1. Flowchart of the approach used for speech emotion recognition.

practically useless even for small feature sets, as the algorithm complexity is $O(2^D)$, where D is the cardinality of the complete feature set. Sequential search algorithms add or remove features one at a time. For example, either starting from an empty set they add incrementally features (*forward*) or starting from the whole set they delete one feature at a time (*backward*), or starting from a randomly chosen subset they add or delete features one at a time. Sequential algorithms are simple to implement and provide results fast, since their complexity is $O(D+(D-1)+(D-2)+\cdots+(D-D_1+1))$, where D_1 is the cardinality of the selected feature set. However, sequential algorithms are frequently trapped at local optima of the criterion function. Random search algorithms start from a randomly selected subset and randomly insert or delete feature sets. The use of randomness helps to escape from local optima. Nevertheless, their performance deteriorates for large feature sets [10]. Since all selection strategies, except the exhaustive search, yield local optima, they are often known as sub-optimum selection algorithms for wrappers. In the following, the term optimum will be used to maintain simplicity. One of the most promising feature selection methods for wrappers is the *sequential floating forward selection algorithm* (SFFS) [11]. The SFFS consists of a forward (insertion) step and a conditional backward (deletion) step that partially avoids the local optima of CCR. In this paper, the execution time will be reduced and the accuracy of SFFS will be improved by theoretically driven modifications of the original algorithm. The execution time is reduced by a preliminary statistical test that helps skipping features, which potentially have no discrimination information. The accuracy is improved by another statistical test, known as tenative test, that selects features that yield a statistically significant improvement of CCR.

A popular method for estimating the CCR of a classifier is the *s*-fold *cross-validation*. In this method, the available data-set is divided into a set used for classifier design (i.e. the training set) and a set used for testing the classifier (i.e. the test set). To focus the discussion on the application examined in this paper, the emotional states of the utterances that belong to the design set are considered known, whereas we pretend that the emotional states of the utterances of the test set are unknown. The classifier estimates the emotional state of the utterances that belong to the test set. From the comparison of the estimated with the actual (ground truth) emotional state of the test utterances, an estimate of CCR is obtained. By repeating this procedure several times, the mean CCR over repetitions is estimated and returned as the CCR estimate, that is referred to as MCCR. The parameter s in s-fold refers to the division of the available data-set into design and test sets. That is, the available data-set is divided into s roughly equal subsets, the samples of the $s-1$ subsets are used to train the classifier, and the samples of the remaining subset are used to estimate CCR during testing. The procedure is repeated for each one of the s subsets in a cyclic fashion and the average CCR over the s repetitions constitutes the MCCR [12]. Burman proposed the *repeated s-fold cross-validation* for model selection, which is simply the s-fold cross-validation repeated many times [13]. The variance of the MCCR estimated by the repeated s-fold cross-validation varies less than that measured by the s-fold cross-validation. Throughout the paper, the repeated s-fold cross-validation is simply denoted as cross-validation, since it is the only cross-validation variant studied. It will be assumed that the number of correctly classified utterances during cross-validation repetitions is a random variable that follows the hypergeometric distribution. Therefore, according to the central limit theorem (CLT), the more realizations of the random variable are obtained, the less varies the MCCR. The large number of repetitions required to obtain an MCCR with a narrow confidence interval prolongs the execution time of a wrapper. We will prove a lemma that uses the variance of the hypergeometric r.v. to find an accurate estimate of the variance of CCR without many needless cross-validation repetitions. By estimating the variance of CCR, the width of the confidence interval of CCR for a certain number of cross-validation repetitions can be predicted. By reversing the problem, if the user selects a fixed confidence interval, the number of cross-validation repetitions is obtained.

The core of the theoretical analysis is not limited to the Bayes classifier within SFFS, but it can be applied to any classifier used in the context of any wrapper. To validate the theoretical results, experiments were conducted for speech emotion recognition. However, the scope of this paper is not limited to this particular application.

The outline of this paper is as follows. In Section 2, we make a theoretical analysis that concludes with Lemma 2, which estimates the variance of the number of correctly classified utterances. Section 3 describes the Bayes classifier. In Section 4, statistical tests employing Lemma 2 are used to improve the speed and the accuracy of SFFS, when the criterion for feature selection is the CCR of the Bayes classifier. In Section 5.1, experiments are conducted in order to demonstrate the benefits of the proposed

estimate versus the standard estimate of the variance of the number of correctly classified utterances. In Section 5.2, the proposed SFFS variant is compared against the standard SFFS for selecting prosody features in order to classify speech into emotional states. In Section 5.3, the number of cross-validation repetitions required for an accurate CCR is plotted as a function of various parameters, such as the number of cross-validation folds and the cardinality of the utterance set. Finally, Section 6, concludes the paper by indicating future research directions.

2. Hypergeometric modeling of the number of correctly classified utterances

The major contribution of this section is Lemma 2, where an accurate estimate of the variance of the number of correctly classified utterances is proposed. It will be demonstrated by experiments in Section 5.1, that the proposed estimate of Lemma 2 is many times more accurate than the standard estimate, i.e. the sample variance. First, the notation that is used hereafter is summarized in Table 1.

The path to arrive at Lemma 2 is Axiom 1 → Axiom 2 → Axiom 3 → Lemma 1 → Lemma 2. The following axiom is the basic premise upon the paper is built.

Axiom 1. Let $\boldsymbol{\kappa}$ be a zero-one r.v. that models the correct classification of an utterance u, when an infinite design set $\mathcal{U}_{\mathcal{D}}$ of utterances, denoted by $\mathcal{U}_{\infty}$, is employed to design the classifier during training, i.e. $\mathcal{U}_{\mathcal{D}} \triangleq \mathcal{U}_{\infty}$, with cardinality $N_{\mathcal{D}} = \infty$. Such a case is depicted in Fig. 2(a). That is, $\boldsymbol{\kappa} = 1$ denotes a correct classification of u, whereas $\boldsymbol{\kappa} = 0$ denotes a wrong classification of u. If $P\{\boldsymbol{\kappa} = 1\} = p$ is the probability of correct classification when the classifier is trained on $\mathcal{U}_{\infty}$, then $\boldsymbol{\kappa}$ is a Bernoulli r.v. with parameter $p \in (0, 1)$.

Table 1
Notation

Notation	Definition
$\mathbf{x}$	Random variable (r.v.)
x	A realization of r.v. $\mathbf{x}$
$\underline{\mathbf{x}}$	Random vector (R.V.)
$\underline{\mathrm{x}}$	A realization of R.V. $\underline{\mathbf{x}}$
$\mathcal{X}$	A set of realizations of an r.v.

A pattern recognition problem with an arbitrary number of classes C and feature vectors of any dimensionality D can be treated as a two-class problem. Class one refers to correctly classified instances, whereas class two refers to erroneously classified instances (e.g. utterances). Axiom 1 implies that p includes information about C and D. Therefore, there is no need to focus the analysis on specific cases of C and D. Axiom 1 implies the following axiom.

Axiom 2. Let $\mathbf{x}$ be an r.v. that models the number of correctly classified utterances, when the classifier is trained with an infinite set of utterances, $\mathcal{U}_{\mathcal{D}} \triangleq \mathcal{U}_{\infty}$, and tested on a finite set $\mathcal{U}_{\mathcal{T}}$ of $N_{\mathcal{T}}$ utterances. This assumption is visualized in Fig. 2(b). Then, $\mathbf{x}$ models the number of correctly classified utterances in $N_{\mathcal{T}}$ independent Bernoulli trials, where the probability of correct classification in each trial is p. Accordingly, $\mathbf{x}$ is a binomial r.v.,

$$P\{\mathbf{x} = \mathrm{x}\} = \binom{N_{\mathcal{T}}}{\mathrm{x}} p^{\mathrm{x}} (1-p)^{1-\mathrm{x}}, \quad \mathrm{x} = 0, 1, \ldots, N_{\mathcal{T}}. \tag{1}$$

A classifier is rarely trained with infinite many utterances. Usually, a finite set of N utterances is only available. Let $\mathcal{U}_A$ denote such a finite set of N available utterances. When the cross-validation framework is used, $\mathcal{U}_A$ is divided into a design set $\mathcal{U}_{\mathcal{D}}$ and a test set $\mathcal{U}_{\mathcal{T}}$, that are disjoint, i.e. $\mathcal{U}_{\mathcal{D}} \cap \mathcal{U}_{\mathcal{T}} = \emptyset$ and $\mathcal{U}_{\mathcal{D}} \cup \mathcal{U}_{\mathcal{T}} = \mathcal{U}_A$, where $N_{\mathcal{D}} + N_{\mathcal{T}} = N$. The procedure is repeated B times, resulting to a set $\mathcal{X} = \{\mathrm{x}_b\}_{b=1}^{B}$ of B realizations of the r.v. $\mathbf{x}$. These conditions imply that $\mathbf{x}$ follows the hypergeometric distribution, as is explained in the following axiom.

Axiom 3. Let $\mathcal{U}_A \subset \mathcal{U}_{\infty}$ be the set of N available utterances, that is divided into disjoint design and test sets, $\mathcal{U}_{\mathcal{D}}$ and $\mathcal{U}_{\mathcal{T}}$, respectively. Such a case is depicted in Fig. 2(c). Let X be the number of correctly classified utterances, when $\mathcal{U}_A$ is used for both training and testing. Then the number of correctly classified utterances $\mathbf{x}$ is an r.v. that follows the hypergeometric distribution with parameters N, $N_{\mathcal{T}}$, and X, i.e.

$$P\{\mathbf{x} = \mathrm{x}\} = \frac{\binom{X}{\mathrm{x}}\binom{N-X}{N_{\mathcal{T}}-\mathrm{x}}}{\binom{N}{N_{\mathcal{T}}}},$$
$$\max(0, X + N_{\mathcal{T}} - N) \leqslant \mathrm{x} \leqslant \min(X, N_{\mathcal{T}}). \tag{2}$$

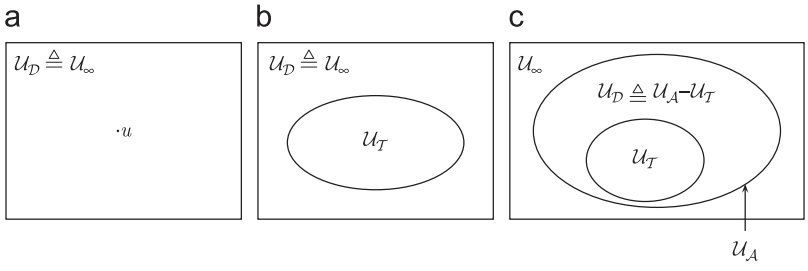

Fig. 2. Visualization of design and test sets of a classifier with: (a) infinite design set and a single test utterance, (b) infinite design set and finite test set, and (c) finite design set and finite test set sampled from an available set of utterances $\mathcal{U}_A$.

Axiom 2 refers to sampling an infinite set, whereas Axiom 3 does sampling a finite set. So, Axiom 3 fits conceptually better to cross-validation, that also performs sampling from a finite set of utterances.

The usual estimate of the variance of the number of correctly classified utterances is the sample dispersion

$$\widehat{\mathrm{Var}}(\mathbf{x}) = \frac{1}{B-1}\sum_{b=1}^{B}(\mathrm{x}_b - \overline{\mathrm{x}}_B)^2. \tag{3}$$

An unbiased estimate of the first moment of $\mathbf{x}$ in cross-validation is the average of x_b during B cross-validation repetitions

$$\overline{\mathrm{x}}_B = \frac{1}{B}\sum_{b=1}^{B}\mathrm{x}_b. \tag{4}$$

A more accurate estimate of $\mathrm{Var}(\mathbf{x})$ than (3) will be derived by Lemma 2 that exploits the hypergeometric modeling of Axiom 3. First, the following lemma directly deduced from Axiom 3 is described.

Lemma 1. *The variance of the number of correctly classified utterances during cross-validation is*

$$\mathrm{Var}(\mathbf{x}) = \frac{N^2}{s^2}\frac{s-1}{N-1}\frac{X}{N}\left(1-\frac{X}{N}\right). \tag{5}$$

Proof. Since $\mathbf{x}$ is a hypergeometric r.v. with parameters $N, N_{\mathscr{T}}$, and X (Axiom 3), then the variance of $\mathbf{x}$ is given by [14]

$$\mathrm{Var}(\mathbf{x}) = \frac{N_{\mathscr{T}}(N-N_{\mathscr{T}})}{(N-1)}\frac{X}{N}\left(1-\frac{X}{N}\right). \tag{6}$$

Given that $N_{\mathscr{T}} = \frac{N}{s}$, (5) results. □

Lemma 2. *An unbiased estimate of the variance of the number of correctly classified utterances during cross-validation is*

$$\widehat{\widehat{\mathrm{Var}}}(\mathbf{x}) = \frac{N^2}{s^2}\frac{s-1}{N-1}\frac{\overline{\mathrm{x}}_B}{N_{\mathscr{T}}}\left(1-\frac{\overline{\mathrm{x}}_B}{N_{\mathscr{T}}}\right). \tag{7}$$

Proof. The first moment of the hypergeometric r.v. is [14]

$$E(\mathbf{x}) = N_{\mathscr{T}}\frac{X}{N}. \tag{8}$$

An unbiased estimate of X, denoted as $\widehat{X}$, can be found by equating (4) with (8),

$$\overline{\mathrm{x}}_B = N_{\mathscr{T}}\frac{\widehat{X}}{N} \Rightarrow \frac{\widehat{X}}{N} = \frac{\overline{\mathrm{x}}_B}{N_{\mathscr{T}}}. \tag{9}$$

By substitution of (9) in (5) the unbiased estimate of the variance of $\mathbf{x}$ (7) is derived. In Section 5.1, it is demonstrated that the gain by using (7) is much higher than using the standard sample dispersion (3) to estimate the variance of the number of correctly classified utterances. □

In the following section, we describe the design of the Bayes classifier, when the probability density function (pdf) of features for each emotional state is modeled by a Gaussian. Result (7) is employed in order to find an estimate of the variance of CCR of the Bayes classifier.

3. Classifier design

Let $\mathscr{W} = \{w_d\}_{d=1}^{D}$ be a feature set comprising D features w_d that are extracted from the set of available utterances $\mathscr{U}_A$. For example, $\mathscr{W}$ can be the set {average energy contour, variance of first formant, length in s,...} [15].

The notation $\mathscr{U}_A$ is extended with superscript $\mathscr{W}$ in order to indicate that $\mathscr{U}_A^{\mathscr{W}} = \{u_i^{\mathscr{W}}\}_{i=1}^{N}$ is the set of N available utterances, out of which the feature set $\mathscr{W}$ is extracted. Each utterance $u_i^{\mathscr{W}} = (\underline{\mathrm{y}}_i^{\mathscr{W}}, c_i)$ is treated as a pattern consisting of a feature vector $\underline{\mathrm{y}}_i^{\mathscr{W}}$ and a label $c_i \in \{1,2,\ldots,C\}$, where C is the total number of classes. Let Ω_c, $c = 1,2,\ldots,C$ be C classes, which in our case refer to emotional states.

A classifier estimates the label of an utterance by processing the feature vector. The CCR is estimated by the cross-validation method, that calculates the mean over $b = \{1,2,\ldots,B\}$ CCRs as follows. Let s be the number of folds the data should be divided into. To find the bth CCR estimate, $N_{\mathscr{D}} = ((s-1)/s)N$ utterances are randomly selected without re-substitution from $\mathscr{U}_A^{\mathscr{W}}$ to build the design set $\mathscr{U}_{\mathscr{D}b}^{\mathscr{W}}$, while the remaining N/s utterances form the test set $\mathscr{U}_{\mathscr{T}b}^{\mathscr{W}}$. This procedure is depicted in Fig. 3. Usually, $s = 5$, 10, or 20 in order to split $\mathscr{U}_A^{\mathscr{W}}$ into design/test sets. In the experiments conducted in Section 5, $s = 5$.

Let us estimate the CCR committed by the *Bayes classifier* in cross-validation repetition b. For utterance $u_i^{\mathscr{W}} = (\underline{\mathrm{y}}_i^{\mathscr{W}}, c_i) \in \mathscr{U}_{\mathscr{T}b}^{\mathscr{W}}$, the class label $\hat{c}_i$ returned by the Bayes classifier is

$$\hat{c}_i = \underset{c=1}{\overset{C}{\operatorname{argmax}}}\{p_b(\underline{\mathrm{y}}_i^{\mathscr{W}}|\Omega_c)P(\Omega_c)\}, \tag{10}$$

where $P(\Omega_c)$ is the a priori class probability which is set to $1/C$, because all emotional states are equiprobable in the data-sets to be used in Section 5, and $p_b(\underline{\mathrm{y}}_i^{\mathscr{W}}|\Omega_c)$ is the class conditional pdf of the feature vector of utterance $u_i^{\mathscr{W}}$ given Ω_c. The class conditional pdf is modeled as a *multivariate Gaussian*, where the mean vector and the covariance matrix are estimated by the sample mean vector and the sample dispersion matrix, respectively.

Let $\mathscr{L}[c_i,\hat{c}_i]$ denote the *zero–one loss function* between the label c_i and the predicted class label $\hat{c}_i$ returned by the Bayes classifier for $u_i^{\mathscr{W}}$, i.e.

$$\mathscr{L}[c_i,\hat{c}_i] = \begin{cases} 1 & \text{if } c_i = \hat{c}_i, \\ 0 & \text{if } c_i \neq \hat{c}_i. \end{cases} \tag{11}$$

Let also $\mathrm{x}_b^{\mathscr{W}}$ be the number of utterances in the test set $\mathscr{U}_{\mathscr{T}b}^{\mathscr{W}} \subset \mathscr{U}_A^{\mathscr{W}}$ that are correctly classified in repetition b, when using feature set $\mathscr{W}$. Then from (11), we have

$$\mathrm{x}_b^{\mathscr{W}} = \sum_{u_i^{\mathscr{W}} \in \mathscr{U}_{\mathscr{T}b}^{\mathscr{W}}} \mathscr{L}[c_i,\hat{c}_i] \tag{12}$$

and the estimate of CCR in repetition b using set $\mathscr{U}_A^{\mathscr{W}}$ is

$$\mathrm{CCR}_b(\mathscr{U}_A^{\mathscr{W}}) = \frac{\mathrm{x}_b^{\mathscr{W}}}{N_{\mathscr{T}}}. \tag{13}$$

The CCR over B cross-validation repetitions is given by

$$\mathrm{MCCR}_B(\mathscr{U}_A^{\mathscr{W}}) = \frac{1}{B}\sum_{b=1}^{B}\mathrm{CCR}_b(\mathscr{U}_A^{\mathscr{W}}), \tag{14}$$

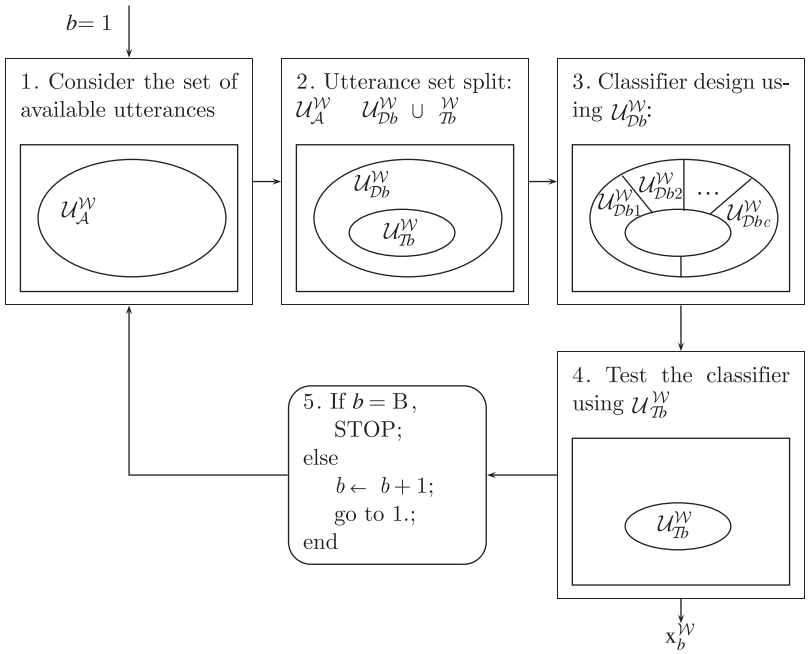

Fig. 3. Cross-validation method to obtain estimates of the correct classification rate $\frac{x_b^{\mathscr{W}}}{N_{\mathscr{T}}}$ for $b = 1, 2, \ldots, B$ repetitions. $\mathscr{U}_{\mathscr{D}bc}^{\mathscr{W}}$ is found by $\{\mathscr{U}_{\mathscr{D}b}^{\mathscr{W}} \cap \Omega_c\}$.

which according to (4) and (3), it is rewritten as

$$\mathrm{MCCR}_B(\mathscr{U}_{\mathscr{A}}^{\mathscr{W}}) = \frac{\overline{x}_B^{\mathscr{W}}}{N_{\mathscr{T}}}. \tag{15}$$

The variance of the correct classification rate (VCCR) estimated from B cross-validation repetitions is given by

$$\widehat{\mathrm{VCCR}}_B(\mathscr{U}_{\mathscr{A}}^{\mathscr{W}}) = \frac{1}{B-1}\sum_{b=1}^{B}[\mathrm{CCR}_b(\mathscr{U}_{\mathscr{A}}^{\mathscr{W}}) - \mathrm{MCCR}_B(\mathscr{U}_{\mathscr{A}}^{\mathscr{W}})]^2. \tag{16}$$

Thus, by substitution of (13) and (15) into (16), and given that $N_{\mathscr{T}} = N/s$, we obtain

$$\widehat{\mathrm{VCCR}}_B(\mathscr{U}_{\mathscr{A}}^{\mathscr{W}}) = \frac{1}{N_{\mathscr{T}}^2}\underbrace{\frac{1}{B-1}\sum_{b=1}^{B}(x_b - \overline{x}_B)^2}_{\widehat{\mathrm{Var}}(\mathbf{x})} = \frac{s^2}{N^2}\widehat{\mathrm{Var}}(\mathbf{x}). \tag{17}$$

According to Lemma 2, another estimate of $\mathrm{Var}(\mathbf{x})$ is (7). So we propose the following estimate of VCCR:

$$\widehat{\widehat{\mathrm{VCCR}}}_B(\mathscr{U}_{\mathscr{A}}^{\mathscr{W}}) = \frac{s^2}{N^2}\widehat{\widehat{\mathrm{Var}}}(\mathbf{x}) = \frac{s-1}{N-1}\frac{\overline{x}_B}{N_{\mathscr{T}}}\left(1 - \frac{\overline{x}_B}{N_{\mathscr{T}}}\right) = \frac{s-1}{N-1}\mathrm{MCCR}_B(\mathscr{U}_{\mathscr{A}}^{\mathscr{W}}) \times (1 - \mathrm{MCCR}_B(\mathscr{U}_{\mathscr{A}}^{\mathscr{W}})). \tag{18}$$

The comparison of $\widehat{\widehat{\mathrm{VCCR}}}_B(\mathscr{U}_{\mathscr{A}}^{\mathscr{W}})$ given by (18) against $\widehat{\mathrm{VCCR}}_B(\mathscr{U}_{\mathscr{A}}^{\mathscr{W}})$ given by (16) for the same number of cross-validation repetitions $B = 10$ is treated in Section 5.1. Next, it is shown that the computational burden of a feature selection method is reduced by using $\widehat{\mathrm{VCCR}}_B(\mathscr{U}_{\mathscr{A}}^{\mathscr{W}})$.

4. Feature selection

The SFFS algorithm finds an optimum subset of features by insertions (i.e. by appending a new feature to the subset of previously selected features) and deletions (i.e. by discarding a feature from the subset of already selected features) as follows. Let m be the counter of insertions and exclusions. Initially ($m = 0$), the subset of selected features $\mathscr{Z}_m$ is the empty set and the maximum CCR achieved is $J(m) = 0$. A total number of M insertions and exclusions are executed in order to find the subset of features that achieves the highest CCR. A typical value for M is 25. However, M is set to 100 for a more detailed study of SFFS in the experiments of Section 5. The SFFS is depicted in Fig. 4.

Feature insertion (*steps* 1 *and* 2): At an insertion step, we seek the feature $w^+ \in \mathscr{W} - \mathscr{Z}_m$ to include in $\mathscr{Z}_m$ such that

$$w^+ = \underset{w \in \mathscr{W} - \mathscr{Z}_m}{\mathrm{argmax}}\, \mathrm{MCCR}_B(\mathscr{U}_{\mathscr{A}}^{\mathscr{Z}_m \cup \{w\}}), \tag{19}$$

where B is the constant number of cross-validation repetitions set by the user. If B is too large, SFFS becomes computational demanding, whereas for small B, $\mathrm{MCCR}_B(\mathscr{U}_{\mathscr{A}}^{\mathscr{Z}_m \cup \{w\}})$ is an inaccurate estimator of the CCR due to the variance of $\mathrm{CCR}_b(\mathscr{U}_{\mathscr{A}}^{\mathscr{Z}_m \cup \{w\}})$. A typical value for B is 50. In Section 4.1, we assume that B is 50. However, a method to estimate B is proposed in Section 4.2. Once w^+ is found, it is included in the subset of selected features $\mathscr{Z}_{m+1} = \mathscr{Z}_m \cup \{w_m^+\}$, the highest CCR is updated $J(m+1) = \mathrm{MCCR}_B(\mathscr{U}_{\mathscr{A}}^{\mathscr{Z}_m \cup \{w^+\}})$, and the counter increases by one $m{:=}m+1$.

Feature deletion (*steps* 4–6): To avoid the local optima, after the insertion of a feature, a conditional deletion step

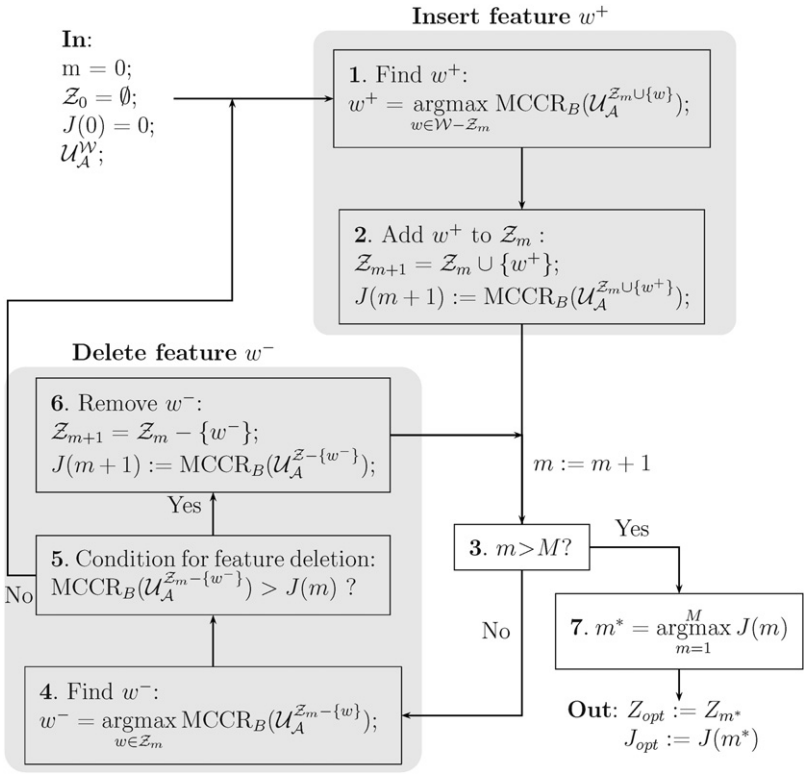

Fig. 4. The sequential floating forward algorithm (SFFS).

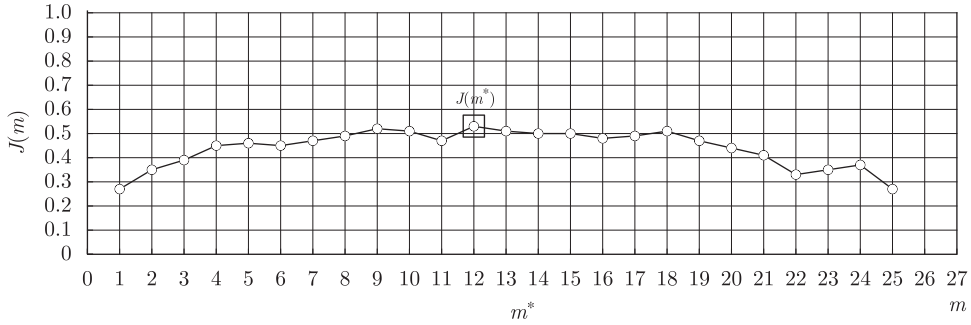

Fig. 5. Plot of $J(m)$ versus the number of feature insertions and deletions m.

is examined. At a deletion step, we seek the feature $w^- \in \mathcal{Z}_m$ such that

$$w^- = \underset{w \in \mathcal{Z}_m}{\text{argmax}}\, \text{MCCR}_B(\mathcal{U}_{\mathcal{A}}^{\mathcal{Z}_m - \{w\}}). \tag{20}$$

If

$$\text{MCCR}_B(\mathcal{U}_{\mathcal{A}}^{\mathcal{Z}_m - \{w^-\}}) > J(m) \tag{21}$$

in Step 5, the deletion of feature w^- from the subset of selected features improves the highest CCR and w^- should be discarded from $\mathcal{Z}_m$ (Step 6). Otherwise, a forward step follows (Step 1). After deleting one particular feature (Step 6), another feature is searched for a deletion (Step 4).

Output procedure (*step* 7 *and out*): After M insertions and deletions having occurred, the algorithm stops. The plot of J versus m, $J(m)$, is examined in order to find out, the specific value of m for which $J(m)$ admits the maximum value, i.e.

$$m^* = \underset{m=1}{\overset{M}{\text{argmax}}} J(m). \tag{22}$$

An example for $M = 25$ is depicted in Fig. 5. It is seen, that the highest $J(m)$ is achieved at $m^* = 12$. Then, the optimum feature subset is selected as $Z_{opt} := Z_{m^*}$ that achieves CCR equal to $J_{opt} = J(m^*)$.

4.1. Method A: how unnecessary comparisons can be avoided during the determination of w^+ in step 1 and w^- in step 4

In this section, we will develop a mechanism that does not allow more than 10 repetitions for feature sets that potentially do not possess any discriminative information. The proposed mechanism is based on the fact that comparisons are done according to confidence limits of CCR instead of the average values of CCR, i.e. the notion of variance of CCR is adopted. In this context, we will employ Lemma 2 to find an accurate estimate of the variance of CCR.

The standard method to determine w^+ in step 1 is pictorially explained in Fig. 6(a). D' candidate features that have not been selected, i.e. $w_1, w_2, \ldots, w_{D'} \in \mathscr{W} - \mathscr{Z}_m$ are sequentially compared in order to find which feature when appended to the subset of previously selected features yields the greatest $\mathrm{MCCR}_{50}(\mathscr{U}_A^{\mathscr{Z}_m \cup \{w_d\}})$. The comparison is done as follows: if candidate feature w_d yields an $\mathrm{MCCR}_{50}(\mathscr{U}_A^{\mathscr{Z}_m \cup \{w_d\}})$ greater than the currently stored value $J_{\mathrm{cur}} = \mathrm{MCCR}_{50}(\mathscr{U}_A^{\mathscr{Z}_m \cup \{w_{\mathrm{cur}}\}})$, where w_{cur} is the best feature to be inserted so far, then J_{cur} is updated with $J_{\mathrm{cur}} := \mathrm{MCCR}_{50}(\mathscr{U}_A^{\mathscr{Z}_m \cup \{w_d\}})$ and $w_{\mathrm{cur}} := w_d$. Otherwise the algorithm proceeds to the next candidate feature w_d. When all the D' candidate features have been examined, the feature to be inserted, w^+, is stored in w_{cur}. In the same manner, w^- can be determined in step 4.

Frequently, $B = 50$ cross-validation repetitions are not necessary to validate whether

$$\mathrm{MCCR}_B(\mathscr{U}_A^{\mathscr{Z}_m \cup \{w_d\}}) < J_{\mathrm{cur}} \tag{23}$$

holds, where $J_{\mathrm{cur}} = \mathrm{MCCR}_{50}(\mathscr{U}_A^{\mathscr{Z}_m \cup \{w_{\mathrm{cur}}\}})$ with w_{cur} being the best candidate for w^+ found so far. Such a case is depicted in Fig. 7. It is seen that $\mathrm{MCCR}_\infty(\mathscr{U}_A^{\mathscr{Z}_m \cup \{w_d\}}) \ll J_{\mathrm{cur}}$. Eq. (23) can be validated with a small number of cross-validation repetitions, e.g. 10, thanks to a statistical test. We propose to formulate a statistical test in order to test the hypothesis H_1, whether (23) holds at 95% significance level for a small number of cross-validation repetitions (e.g. 10). H_1 is accepted if

$$\frac{k^{\mathrm{u}}_{10;a;w_d}}{N_{\mathscr{T}}} < J_{\mathrm{cur}}, \tag{24}$$

where $k^{\mathrm{u}}_{10;a;w_d}$ is the upper confidence limit of $\mathbf{x}^{\mathscr{Z}_m \cup \{w_d\}}$ (which is a hypergeometric r.v. according to Axiom 3) for $B = 10$ cross-validation repetitions and $a = 0.95$ implies the $100a\%$ level of significance. If (24) is valid, then

$$\mathrm{MCCR}_\infty(\mathscr{U}_A^{\mathscr{Z}_m \cup \{w_d\}}) < \frac{k^{\mathrm{u}}_{10;a;w_d}}{N_{\mathscr{T}}} < J_{\mathrm{cur}} \tag{25}$$

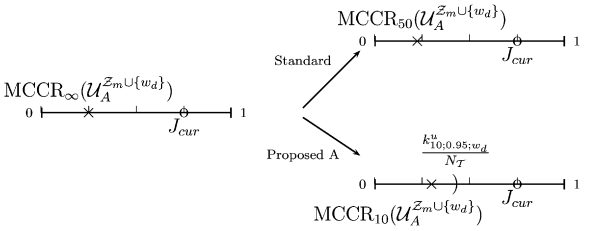

Fig. 7. Visualization of the proposed method A on CCR axis for case H_1: w^d should be preliminarily excluded.

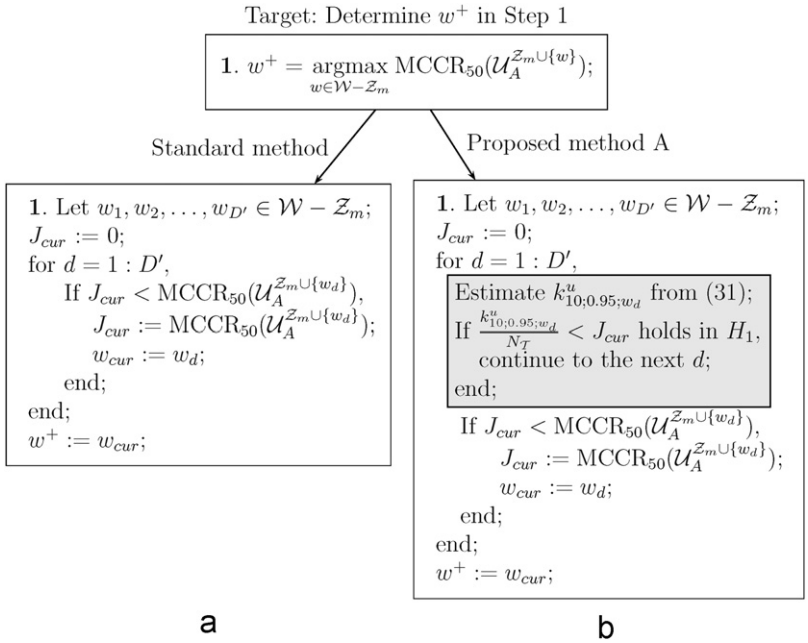

Fig. 6. Comparison between the standard method versus the proposed method A for implementing step 1. The lines in the gray box constitute the proposed preliminary test in order to exclude a feature with only 10 cross-validation repetitions. If it is not excluded, then $B = 50$ cross-validation repetitions are executed in order to make a more thorough evaluation.

will be also valid, because $k^{\mathrm{u}}_{10;a;w_d}/N_{\mathscr{F}}$ is the upper confidence limit of $\mathrm{MCCR}_{\infty}(\mathscr{U}_A^{\mathscr{F}_m\cup\{w_d\}})$. So (23) is validated with $B=10$ repetitions instead of $B=50$ repetitions.

If (24) is not valid, two cases are possible, namely,

$$H_2:\quad \mathrm{MCCR}_{\infty}(\mathscr{U}_A^{\mathscr{F}_m\cup\{w_d\}})<J_{\mathrm{cur}}<\frac{k^{\mathrm{u}}_{10;a;w_d}}{N_{\mathscr{F}}}. \tag{26}$$

In this case, 50 repetitions (instead of ∞) are conducted in order to tentatively exclude w_d. The last case is

$$H_3:\quad J_{\mathrm{cur}}<\mathrm{MCCR}_{\infty}(\mathscr{U}_A^{\mathscr{F}_m\cup\{w_d\}})<\frac{k^{\mathrm{u}}_{10;a;w_d}}{N_{\mathscr{F}}}, \tag{27}$$

and 50 repetitions (instead of ∞) should be conducted to tentatively accept w_d as w_{cur}.

In (24), $k^{\mathrm{u}}_{10;a;w_d}$ can be estimated by two methods. That is, either by an approximation with the upper confidence limit of a Gaussian r.v. or by the summation of a discrete hypergeometric pdf. To choose between the two methods, the prerequisite $\mathrm{Var}(\mathbf{x}^{\mathscr{F}_m\cup\{w_d\}})>9$ is used as a switch [16], where the estimate of $\mathrm{Var}(\mathbf{x}^{\mathscr{F}_m\cup\{w_d\}})$ can be obtained by (7)

$$\frac{N^2}{s^2}\frac{s-1}{N-1}\frac{\overline{\mathrm{x}}_B^{\mathscr{F}_m\cup\{w_d\}}}{N_{\mathscr{F}}}\left(1-\frac{\overline{\mathrm{x}}_B^{\mathscr{F}_m\cup\{w_d\}}}{N_{\mathscr{F}}}\right)>9, \tag{28}$$

where N is the total number of utterances, $N_{\mathscr{F}}=N/s$, and $\overline{\mathrm{x}}_B^{\mathscr{F}_m}$ is the average over B realizations of $\mathbf{x}^{\mathscr{F}_m}$. By invoking (15), (28) becomes

$$\frac{N^2}{s^2}\frac{s-1}{N-1}\mathrm{MCCR}_B(\mathscr{U}_A^{\mathscr{F}_m\cup\{w_d\}}) \times(1-\mathrm{MCCR}_B(\mathscr{U}_A^{\mathscr{F}_m\cup\{w_d\}}))>9. \tag{29}$$

First, if (29) holds, the hypergeometric r.v. $\mathrm{x}_b^{\mathscr{F}_m\cup\{w_d\}}$ is approximated by a Gaussian one, and $k^{\mathrm{u}}_{10;a;w_d}$ can be found by

$$k^{\mathrm{u}}_{10;a;w_d}=\overline{\mathrm{x}}_{10}^{\mathscr{F}_m\cup\{w_d\}}+z_a\sqrt{\frac{\mathrm{Var}(\mathbf{x}^{\mathscr{F}_m\cup\{w_d\}})}{10}} \tag{30}$$

where z_a equals 1.96 for $a=0.95$. If (7) is used, then

$$k^{\mathrm{u}}_{10;a;w_d}=\overline{\mathrm{x}}_{10}^{\mathscr{F}_m\cup\{w_d\}} +z_a\sqrt{\frac{N^2}{s^2}\frac{s-1}{N-1}\frac{\overline{\mathrm{x}}_{10}^{\mathscr{F}_m\cup\{w_d\}}}{N_{\mathscr{F}}}\left(1-\frac{\overline{\mathrm{x}}_{10}^{\mathscr{F}_m\cup\{w_d\}}}{N_{\mathscr{F}}}\right)\frac{1}{10}}. \tag{31}$$

Second, when (29) is violated, then the confidence limit $k^{\mathrm{u}}_{10;a;w_d}$ is estimated by

$$k^{\mathrm{u}}_{10;a;w_d}=\underset{k_1=0,1,\ldots,N_{\mathscr{F}}}{\mathrm{argmin}}\left|\sum_{k=0}^{k_1}\frac{\binom{\widehat{X}}{k}\binom{N-\widehat{X}}{N_{\mathscr{F}}-k}}{\binom{N}{N_{\mathscr{F}}}}-a\right|, \tag{32}$$

where $\widehat{X}$ according to (9) can be estimated as $\widehat{X}=\overline{\mathrm{x}}_{10}^{\mathscr{F}_m\cup\{w_d\}}(N/N_{\mathscr{F}})$. The proposed method A is depicted in Fig. 6(b). The same mechanism can be applied to speed up step 4 that finds w^-. $B=50$ repetitions may or may not be enough for estimating accurately the MCCR in cases H_2 and H_3. These cases are treated next.

4.2. Method B: increasing the accuracy of accepting a feature w_d as w_{cur}

The proposed method A was focused on case H_1. In this section, the proposed method B is focused on cases H_2 and H_3. Method B is invoked when H_1 is rejected. Then either H_2 or H_3 should be valid. If H_3 is not valid then, by reductio ad absurdum, case H_2 is valid. So, method B should check if H_3 is valid or not by validating if

$$H_3:\quad J_{\mathrm{cur}}<\mathrm{MCCR}_{50}(\mathscr{U}_A^{\mathscr{F}_m\cup\{w_d\}}), \tag{33}$$

where J_{cur} is the greatest CCR achieved so far by the current best candidate for insertion w_{cur}, i.e.

$$J_{\mathrm{cur}}=\mathrm{MCCR}_{50}(\mathscr{U}_A^{\mathscr{F}_m\cup\{w_{\mathrm{cur}}\}}). \tag{34}$$

From now on, we shall replace 50 with an arbitrary number of cross-validation repetitions B. So, cases H_2 and H_3 are defined as follows:

$$H_2:\quad \mathrm{MCCR}_{\infty}(\mathscr{U}_A^{\mathscr{F}_m\cup\{w_{\mathrm{cur}}\}})\geqslant\mathrm{MCCR}_{\infty}(\mathscr{U}_A^{\mathscr{F}_m\cup\{w_d\}}), \tag{35}$$

$$H_3:\quad \mathrm{MCCR}_{\infty}(\mathscr{U}_A^{\mathscr{F}_m\cup\{w_{\mathrm{cur}}\}})<\mathrm{MCCR}_{\infty}(\mathscr{U}_A^{\mathscr{F}_m\cup\{w_d\}}). \tag{36}$$

w_{cur} should be updated by w_d only in the case H_3. Otherwise the candidate w_d deteriorates or does not improve MCCR more than w_{cur} does (H_2).

Let

$$\left[\frac{k^{\mathrm{l}}_{B;a;w_{\mathrm{cur}}}}{N_{\mathscr{F}}},\frac{k^{\mathrm{u}}_{B;a;w_{\mathrm{cur}}}}{N_{\mathscr{F}}}\right] \tag{37}$$

be the confidence interval of $\mathrm{MCCR}_B(\mathscr{U}_A^{\mathscr{F}_m\cup\{w_{\mathrm{cur}}\}})$ at $a=95\%$ confidence level, and accordingly let

$$\left[\frac{k^{\mathrm{l}}_{B;a;w_d}}{N_{\mathscr{F}}},\frac{k^{\mathrm{u}}_{B;a;w_d}}{N_{\mathscr{F}}}\right] \tag{38}$$

be the confidence interval of $\mathrm{MCCR}_B(\mathscr{U}_A^{\mathscr{F}_m\cup\{w_d\}})$. In order to validate whether (36) holds at 100a% confidence level, the lower confidence limit of the right part of (36) should be greater than the upper limit of the left part, i.e.

$$\frac{k^{\mathrm{l}}_{B;a;w_d}}{N_{\mathscr{F}}}>\frac{k^{\mathrm{u}}_{B;a;w_{\mathrm{cur}}}}{N_{\mathscr{F}}}. \tag{39}$$

In the feature selection experiments, upper and lower confidence intervals are symmetrical around the mean, because the normality assumption (29) is fulfilled for values $N>360$, $N_{\mathscr{F}}=N/5$, and $0.2<\mathrm{MCCR}_B(\mathscr{U}_A^{\mathscr{F}_m})<0.6$. Thus according to (30)

$$\left[\frac{k^{\mathrm{l}}_{B;a;w_d}}{N_{\mathscr{F}}},\frac{k^{\mathrm{u}}_{B;a;w_d}}{N_{\mathscr{F}}}\right]=\left[\frac{\overline{\mathrm{x}}_B^{\mathscr{F}_m\cup\{w_d\}}}{N_{\mathscr{F}}}-\frac{z_a}{N_{\mathscr{F}}}\sqrt{\frac{\mathrm{Var}(\mathbf{x}^{\mathscr{F}_m\cup\{w_d\}})}{B}},\ \frac{\overline{\mathrm{x}}_B^{\mathscr{F}_m\cup\{w_d\}}}{N_{\mathscr{F}}}+\frac{z_a}{N_{\mathscr{F}}}\sqrt{\frac{\mathrm{Var}(\mathbf{x}^{\mathscr{F}_m\cup\{w_d\}})}{B}}\right]. \tag{40}$$

The confidence interval for (37) can be derived similarly.

It is a common practice in statistics that the confidence interval of the expected value of an r.v. to have a fixed width [17, exercise 8.13]. Such a fixed width confidence interval is derived by employing the CLT. That is, the more the repetitions B of the experiment are, the smaller is the confidence interval width of the average of the r.v. In our case, we wish to find an estimate of B denoted as $\hat{B}$ so that

the width of the confidence interval (40) for any w_d is fixed

$$\frac{k^{\mathrm{u}}_{\hat{B};a;w_d}}{N_{\mathscr{F}}} - \frac{k^{\mathrm{l}}_{\hat{B};a;w_d}}{N_{\mathscr{F}}} = \gamma \tag{41}$$

(e.g. $\gamma = 1.25\%$), which according to (40) is equivalent to

$$\frac{\overline{\mathrm{x}}^{\mathscr{F}_m\cup\{w_d\}}_{\hat{B}}}{N_{\mathscr{F}}} + \underbrace{\frac{z_a}{N_{\mathscr{F}}}\sqrt{\frac{\mathrm{Var}(\mathbf{x}^{\mathscr{F}_m\cup\{w_d\}})}{\hat{B}}}}_{\gamma/2} - \left(\frac{\overline{\mathrm{x}}^{\mathscr{F}_m\cup\{w_d\}}_{\hat{B}}}{N_{\mathscr{F}}} - \underbrace{\frac{z_a}{N_{\mathscr{F}}}\sqrt{\frac{\mathrm{Var}(\mathbf{x}^{\mathscr{F}_m\cup\{w_d\}})}{\hat{B}}}}_{\gamma/2}\right) = \gamma, \tag{42}$$

that yields

$$\hat{B} = \frac{4z_a^2\mathrm{Var}(\mathbf{x}^{\mathscr{F}_m\cup\{w_d\}})}{N_{\mathscr{F}}^2\gamma^2}. \tag{43}$$

$\mathrm{Var}(\mathbf{x}^{\mathscr{F}_m\cup\{w_d\}})$ can be estimated by (7) for 10 repetitions. By using also the fact that $N_{\mathscr{F}} = N/s$, (43) becomes

$$\hat{B}_1 = \frac{4z_a^2}{\gamma^2}\frac{s-1}{N-1}\mathrm{MCCR}_{10}(\mathscr{U}_A^{\mathscr{F}_m\cup\{w_d\}}) \times(1-\mathrm{MCCR}_{10}(\mathscr{U}_A^{\mathscr{F}_m\cup\{w_d\}})) \tag{44}$$

and for the current feature set $\mathscr{F}_m\cup\{w_{\mathrm{cur}}\}$ we obtain similarly

$$\hat{B}_2 = \frac{4z_a^2}{\gamma^2}\frac{s-1}{N-1}\mathrm{MCCR}_{10}(\mathscr{U}_A^{\mathscr{F}_m\cup\{w_{\mathrm{cur}}\}}) \times(1-\mathrm{MCCR}_{10}(\mathscr{U}_A^{\mathscr{F}_m\cup\{w_{\mathrm{cur}}\}})). \tag{45}$$

The user selects γ with respect to the computation load one may afford, as it can be inferred from (44) and (45). Consequently (39) holds if

$$\frac{\overline{\mathrm{x}}^{\mathscr{F}_m\cup\{w_d\}}_{\hat{B}_1}}{N_{\mathscr{F}}} - \frac{\gamma}{2} > \frac{\overline{\mathrm{x}}^{\mathscr{F}_m\cup\{w_{\mathrm{cur}}\}}_{\hat{B}_2}}{N_{\mathscr{F}}} + \frac{\gamma}{2} \tag{46}$$

or equivalently

$$\mathrm{MCCR}_{\hat{B}_2}(\mathscr{U}_A^{\mathscr{F}_m\cup\{w_d\}}) - \mathrm{MCCR}_{\hat{B}_1}(\mathscr{U}_A^{\mathscr{F}_m\cup\{w_{\mathrm{cur}}\}}) > \gamma. \tag{47}$$

In the experiments described in Section 5, γ is set to 0.0125. That is, the feature set $\mathscr{F}_m\cup\{w_d\}$ performs better than the current set $\mathscr{F}_m\cup\{w_{\mathrm{cur}}\}$, if the difference between cross-validated CCR is at least 0.0125. The combination of the proposed method B and the proposed method A is referred to as method AB and depicted in Fig. 8. Since the algorithm is recursive ($w_{\mathrm{cur}}:=w_d$), there is no need to use $\hat{B}_1$ and $\hat{B}_2$, as it was done for the explanation reasons, but just a single variable $\hat{B}$ is sufficient. In Section 5.3, $\hat{B}$ is plotted versus the parameters it depends on according to (44).

An example for cases H_2 and H_3 is visualized in Fig. 9. Two trials per case are allowed. It is seen that the decision taken by the proposed method B, coincides with the ground truth decision for both trials in H_2 as well as H_3, whereas the decision taken by the standard method coincides with the ground truth only for 1 out of the 2 trials.

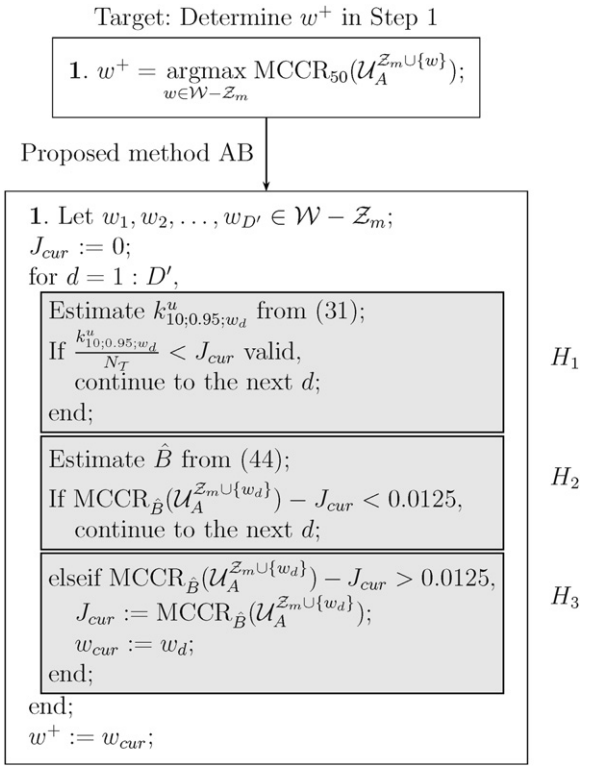

Fig. 8. Combination of the proposed methods A and B to handle preliminarily rejection from the beginning (H_1); tentatively rejection (H_2); and tentative acceptance (H_3) of a candidate feature $\{w_d\}$.

5. Experimental results

The experiments are divided into three parts:

- In Section 5.1, it is demonstrated that the variance estimate proposed by Lemma 2 is more accurate than the sample dispersion.
- In Section 5.2, it is shown that the proposed methods A and B, employing the result of Lemma 2, improve the speed and the accuracy of SFFS for speech emotion recognition.
- In Section 5.3, the number of cross-validation repetitions $\hat{B}$ found by (44) is plotted as a function of the parameters it depends on.

5.1. Comparison of estimates of the variance of CCR

In this section, we shall demonstrate that the proposed estimate of variance of CCR given by (18), i.e

$$\widehat{\mathrm{VCCR}}_{10}(\mathscr{U}_A^{\mathscr{F}}) = \frac{1}{N_{\mathscr{F}}^2}\underbrace{\mathrm{Var}(\mathbf{x}^{\mathscr{F}})}_{\text{Lemma 2}} = \frac{s-1}{N-1}\mathrm{MCCR}_{10}(\mathscr{U}_A^{\mathscr{F}})(1-\mathrm{MCCR}_{10}(\mathscr{U}_A^{\mathscr{F}})) \tag{48}$$

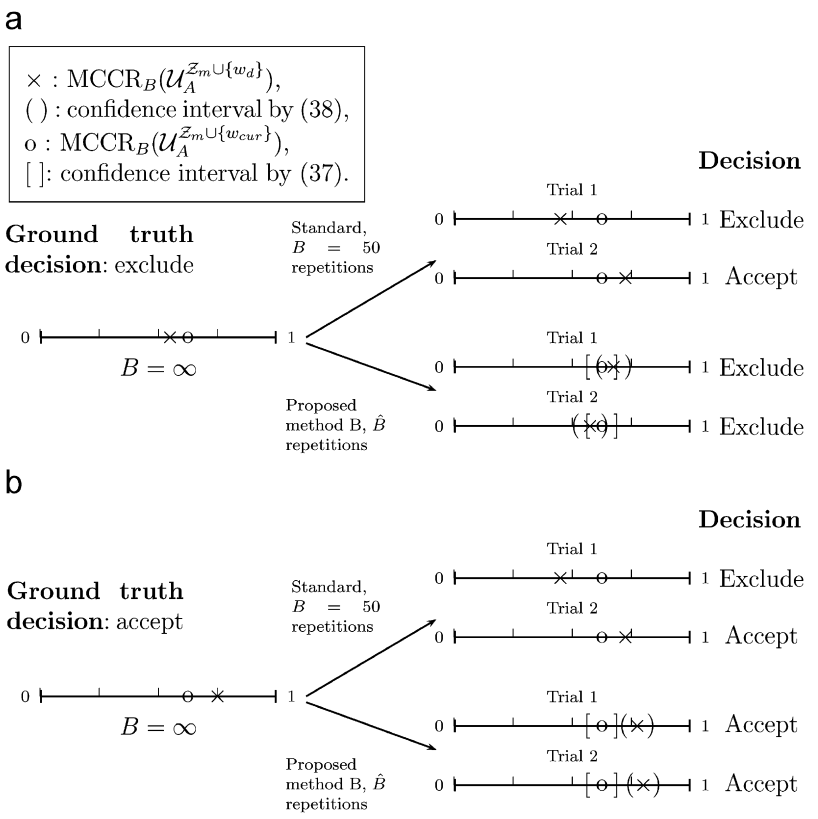

Fig. 9. Comparison between the standard method and the proposed method B in cases H_2 and H_3 on the axis of CCR, when 2 trials are allowed for simplicity. (a) H_2: MCCR$_\infty(\mathcal{U}_A^{\mathcal{Z}_m\cup\{w_{cur}\}}) \geqslant$ MCCR$_\infty(\mathcal{U}_A^{\mathcal{Z}_m\cup\{w_d\}})$, MCCR$_\infty(\mathcal{U}_A^{\mathcal{Z}_m\cup\{w_{cur}\}}) \geqslant$ MCCR$_\infty(\mathcal{U}_A^{\mathcal{Z}_m\cup\{w_d\}})$, (b) H_3: MCCR$_\infty(\mathcal{U}_A^{\mathcal{Z}_m\cup\{w_{cur}\}}) <$ MCCR$_\infty(\mathcal{U}_A^{\mathcal{Z}_m\cup\{w_d\}})$.

is more accurate than the sample dispersion (16) for the same number of repetitions $B = 10$,

$$\widehat{\mathrm{VCCR}}_{10}(\mathcal{U}_A^{\mathscr{Z}}) = \frac{1}{10-1}\sum_{b=1}^{10}[\mathrm{CCR}_b(\mathcal{U}_A^{\mathscr{Z}}) - \mathrm{MCCR}_{10}(\mathcal{U}_A^{\mathscr{Z}})]^2, \quad (49)$$

where

$$\mathrm{MCCR}_{10}(\mathcal{U}_A^{\mathscr{Z}}) = \frac{1}{10}\sum_{b=1}^{10}\mathrm{CCR}_b(\mathcal{U}_A^{\mathscr{Z}}). \quad (50)$$

Among (48) and (49), the most accurate estimate is that being closer to the sample dispersion for an infinite number of repetitions, which it is estimated by 1000 repetitions, i.e.

$$\mathrm{VCCR}_{1000}(\mathcal{U}_A^{\mathscr{Z}}) = \frac{1}{1000-1}\sum_{b=1}^{1000}[\mathrm{CCR}_b(\mathcal{U}_A^{\mathscr{Z}}) - \mathrm{MCCR}_{1000}(\mathcal{U}_A^{\mathscr{Z}})]^2. \quad (51)$$

We have used a few number of repetitions, i.e. $B = 10$, in order to demonstrate that the gain of the proposed estimate (48) against the standard one (49) is great even for a few number of realizations of the r.v. $\mathbf{x}^{\mathscr{Z}}$. It should be reminded that the r.v. $\mathbf{x}^{\mathscr{Z}}$ models the number of correctly classified utterances during cross-validation repetitions.

The experiments are conducted for artificial and real data-sets, and for different selections of parameters N, s, and MCCR. The results are shown in Fig. 10 that consists of six sub-figures. The first row corresponds to experiments with artificially generated data-sets, whereas the second row corresponds to experiments with real data-sets. In each column, two of the three parameters $N, s,$ MCCR are kept constant, whereas the last one varies. In experiments with artificially generated data-sets with C classes, the samples in each class are generated with a multivariate Gaussian random number generator [18]. It should be noted that N denotes the number of samples and C the number of classes, when the discussion refers to the artificially generated data-sets. In experiments, with real data-sets, the utterances stem from two emotional speech data-sets, namely the Danish emotional speech (DES) corpus [19] and speech under simulated and actual stress (SUSAS) corpus [20]. DES consists of $N = 1160$ utterances expressed by four actors under $C = 5$ emotional states, such as *anger*, *happiness*, *neutral*, *sadness*, and *surprise*. SUSAS speech corpus includes $N = 5042$ speech utterances expressed under $C = 8$ styles such as *anger*, *clear*, *fast*, *loud*, *question*, *slow*, *soft*, and *neutral*. Data from nine speakers with three regional accents (i.e. that of Boston, General, and New York) are exploited. Ninety features are

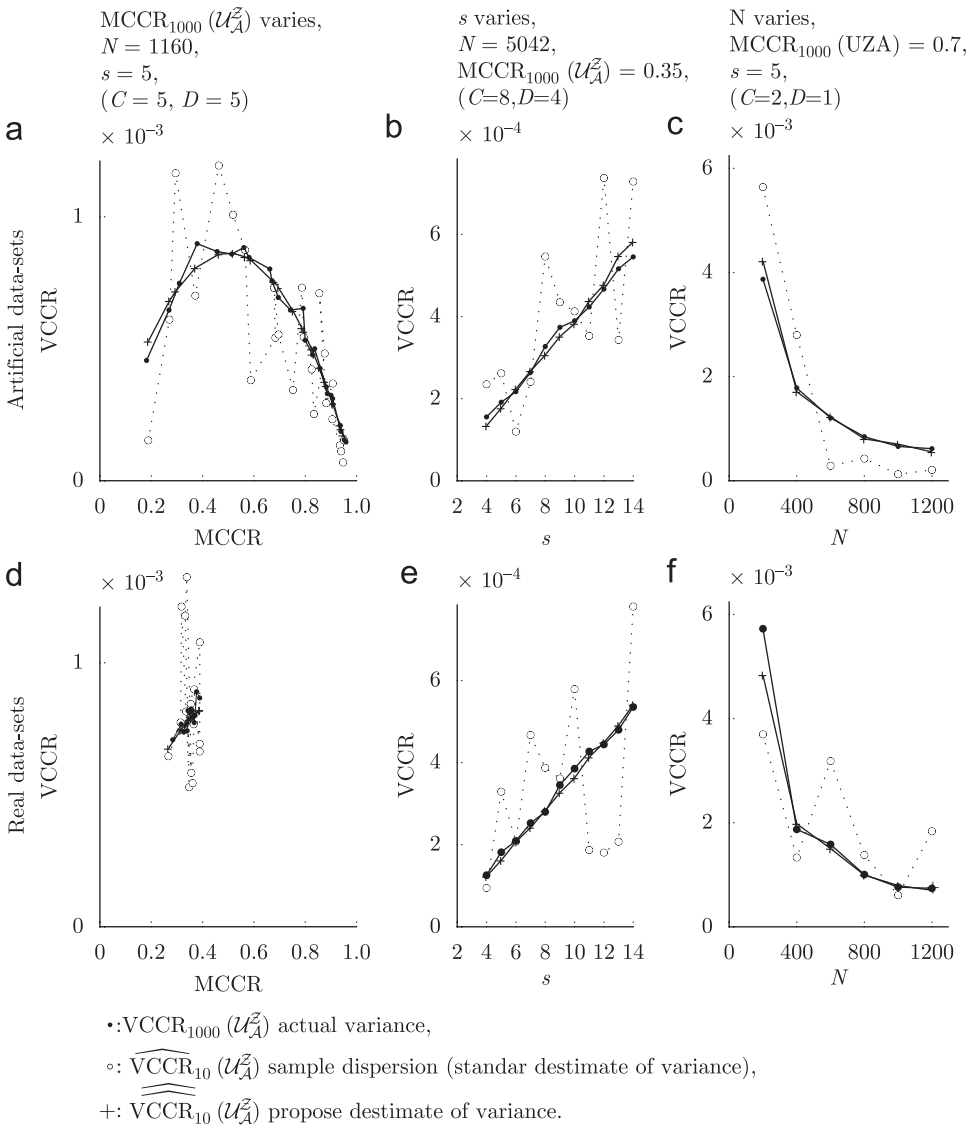

Fig. 10. Proposed estimate (48) against sample dispersion (49) for finding the variance of CCR (51) plotted versus the factors it depends on. (a)–(c) Artificial data sets, (d) DES, (e) SUSAS, and (f) SUSAS (anger vs. neutral).

extracted from the utterances that include the variance, the mean, and the median of pitch, formants, and energy contours [15]. D features are selected from the 90 ones. In Fig. 10(d), 10 randomly chosen subsets $\mathscr{Z}$ of $D = 5$ features out of the whole set $\mathscr{W}$ of 90 features are built. For example, one such feature set comprises the mean duration of the rising slopes of pitch contour, the mean energy value within falling slopes of the energy contour, the energy below 250 Hz, the energy in the frequency band 3500–3950 Hz, and the energy in the frequency band 600–1000 Hz. As it is seen in Fig. 10(d), it is not feasible to achieve a high CCR by using real feature sets on DES. In Fig. 10(e), a feature set of cardinality $D = 4$ is used that comprises the maximum duration of plateaux at maxima, the median of the energy values within the plateaux at maxima, the median duration of the falling slopes of energy contour, and the energy below 600 Hz extracted from database SUSAS. In Fig. 10(f), the single feature employed is the interquartile range of energy value. A different number of classes C and feature dimensionalities D were used in each figure, in order to demonstrate that (48) does not depend on the number of classes and the dimensionality of the feature vector, because the information about C and D is captured by the MCCR parameter.

From the inspection of Figs. 10(a)–(f), it can be inferred that $\widehat{\widehat{\mathrm{VCCR}}}_{10}(\mathscr{U}_A^{\mathscr{Z}})$ is closer to $\mathrm{VCCR}_{1000}(\mathscr{U}_A^{\mathscr{Z}})$ than $\widehat{\mathrm{VCCR}}_{10}(\mathscr{U}_A^{\mathscr{Z}})$ is. Therefore, $\widehat{\widehat{\mathrm{VCCR}}}_{10}(\mathscr{U}_A^{\mathscr{Z}})$ is a more accurate estimator of $\mathrm{VCCR}_{1000}(\mathscr{U}_A^{\mathscr{Z}})$ than $\widehat{\mathrm{VCCR}}_{10}(\mathscr{U}_A^{\mathscr{Z}})$ is.

Next, we shall exploit (48) in order to find an accurate estimate of the variance of the CCR during feature selection.

5.2. Feature selection results

The objective in this section is to demonstrate the improvement in speed and accuracy of the SFFS algorithm, when steps 1 and 4 are implemented with the proposed methods A and B instead of the standard method.

Three data-sets were used: (a) DES with $N=1160$ utterances and $C=5$ classes, (b) a subset of DES with $N=360$ utterances and $C=5$ classes, and (c) SUSAS with $N=5042$ utterances and $C=8$ classes. Each data-set is split into $s=5$ equal subsets; four out of the five subsets are used for training the classifier, whereas the last one is used to test it. The threshold for the total number of insertions and deletions M is equal to 100 in all methods. Execution time for all methods and data-sets is depicted in Fig. 11.

It is seen that the proposed method A reduces the execution time compared to the standard method by 50%. This is due to the fact that standard method performs $B=50$ cross-validation repetitions for all candidate features, whereas the proposed method A performs only 10 repetitions during a preliminary evaluation of features, and if need, another 40 repetitions for a more thorough evaluation. The proposed method AB is slower than the standard method for the DES full-set and the DES subset. This is due to the fact that estimated $\hat{B}=75$ and 140 for the DES full-set and the DES subset, respectively, are greater than $B=50$ of the standard method. For SUSAS data-set, $\hat{B}=20$, and accordingly the proposed AB method is faster than the standard one.

The benefit of the proposed method AB against the other methods is its accuracy. A fact which is addressed next. In Fig. 12, the MCCR curve and its confidence interval with respect to the index of insertion and deletion m are plotted for each method and each data-set. The confidence interval is approximated by (40). For the standard and the proposed method A, the variance in the approximation is estimated by (49), whereas for the proposed method AB the variance is estimated by the proposed estimate (48). Three observations can be made.

First, the maxima of MCCR curves are not affected by the proposed A or the proposed AB methods. A deterioration of MCCR might be claimed for the proposed method AB on DES in Fig. 12(c), but the confidence intervals of Fig. 12(c) overlap with those in Fig. 12(a).

Second, the MCCR curve versus m for the proposed AB method on the DES full-set and DES subset (Figs. 12(c) and (f), respectively) has a clear peak. This fact allows one to select the best subset of features. This was happened, because the confidence intervals are taken into consideration to insert or delete a feature, whereas in the standard method and method A the decision to insert or delete a feature is taken by using only MCCR values from 50 repetitions. MCCR estimates from 50 repetitions are not reliable, because they have a wide confidence interval, as it is seen in Figs. 12(a), (b), (d), and (e). So, the proposed method AB takes more accurate decisions, and therefore, the peak of MCCR is more prominent. The proposed AB method on SUSAS (Fig. 12(i)) does not present a prominent maximum. SUSAS data-set consists of many utterances ($N=5042$), and therefore, the 'curse of dimensionality' effect is not obvious. The peak of MCCR curve will be prominent for $M>100$ and $D>90$.

Third, it is seen that proposed AB method finds fixed confidence intervals for all data-sets, whereas the confidence intervals of the standard and the proposed A methods vary significantly among the data-sets. The greatest width in confidence intervals appears for the DES subset that consists of $N=360$ utterances (Figs. 12(d) and (e)). This confirms (48), where the variance of CCR is inversely proportional to the number of samples N. By selecting an appropriate $\hat{B}$, the proposed AB method finds fixed width confidence intervals for all data-sets.

5.3. The number of cross-validation repetitions $\hat{B}$ plotted as a function of the parameters it depends on

In this section, $\hat{B}$ given by (44) is plotted at $a=95\%$ level of significance, for varying number of folders ($s=2,5$, and 10) the set of utterances $\mathscr{U}^{\mathscr{W}}$ is divided into, varying width of the confidence interval of CCR ($0.0125\leqslant\gamma\leqslant0.05$), different cardinalities of $\mathscr{U}$ ($N=250$, 1000, and 5000), and all possible CCRs found for 10 repetitions, i.e. $0\leqslant \mathrm{MCCR}_{10}(\mathscr{U}^{\mathscr{W}})\leqslant 1$.

A three-dimensional plot of B with respect to γ and $\mathrm{MCCR}_{10}(\mathscr{U}^{\mathscr{W}})$ is shown in Fig. 13(a). N and s are set to 1000 and 10, respectively. It is observed that for $\gamma>0.05$, B is almost 0, whereas as $\gamma\to 0$, then $B\to\infty$. The maximum B for a certain γ is observed for $\mathrm{MCCR}_{10}(\mathscr{U}^{\mathscr{W}})=0.5$. This maximum is shown with a black thick line.

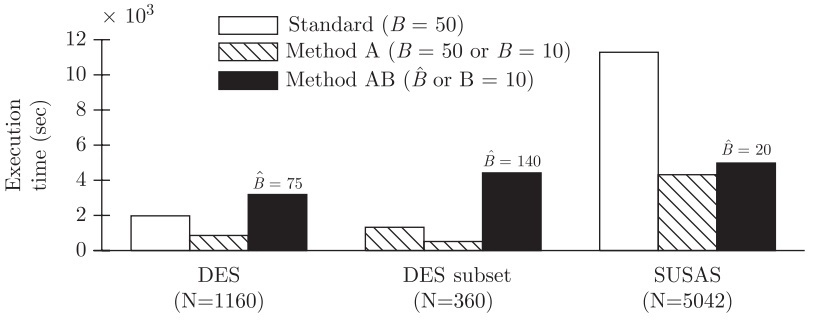

Fig. 11. Execution time of SFFS when using the standard method versus the proposed methods.

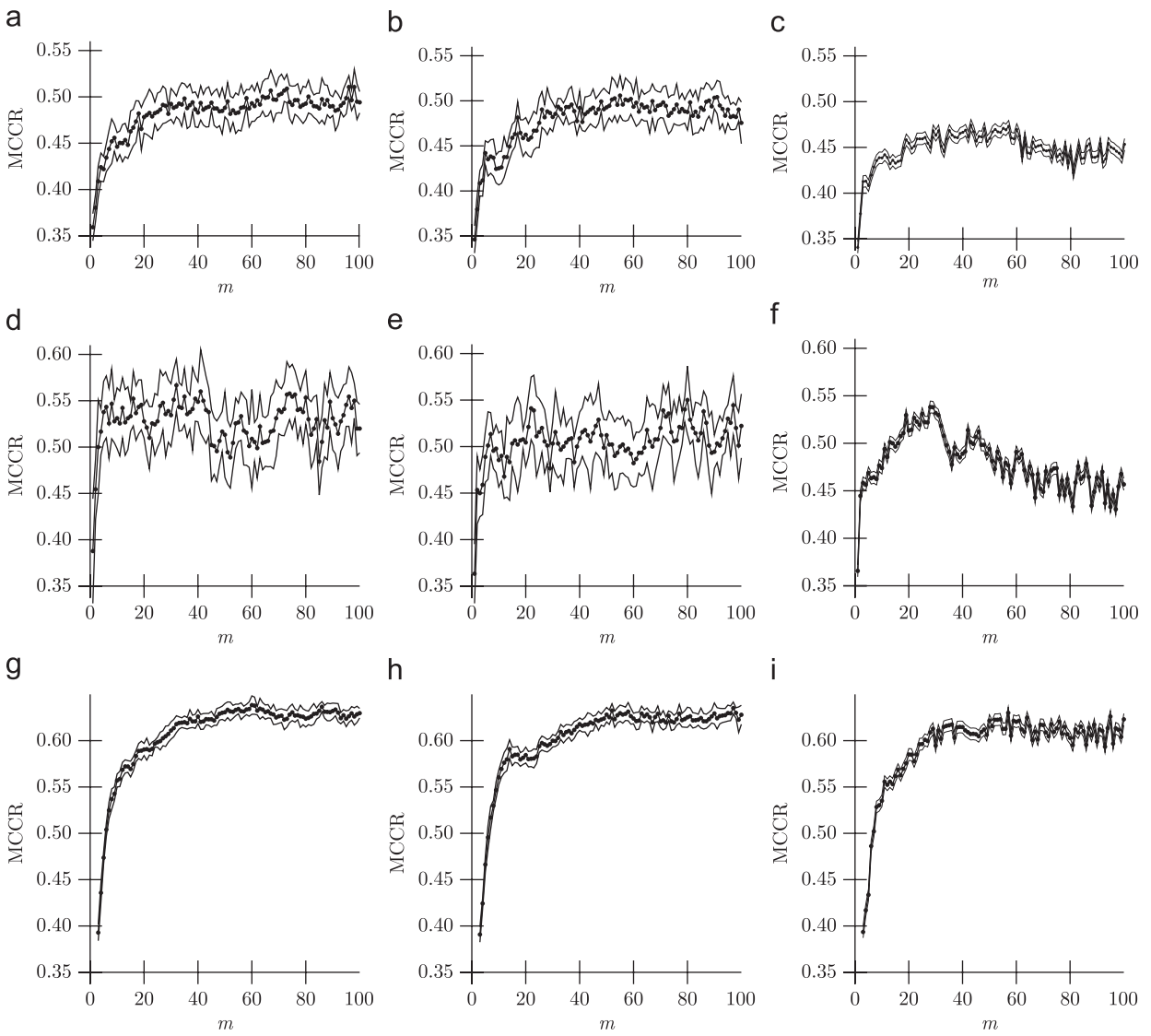

Fig. 12. CCR achieved by standard SFFS and variants with respect to the index of feature inclusion or exclusion m. (a) Standard on DES (N = 1160); (b) Proposed A on DES (N = 1160); (c) Proposed AB on DES (N = 1160); (d) Standard on DES subset (N = 360); (e) Proposed A on DES subset (N = 360); (f) Proposed AB on DES subset (N = 360); (g) Standard on SUSAS (N = 5042); (h) Proposed A on SUSAS (N = 5042); (i) Proposed AB on SUSAS (N = 5042).

In Figs. 13(b)–(d), the same curve is plotted for various values of s and N. B is great when $N = 250$ and $s = 10$, as shown in Fig. 13(b), whereas B is small when $N = 5000$ and $s = 2$, as depicted in Fig. 13(d).

6. Conclusions

In this work, the execution time and the accuracy of SFFS method are optimized by exploiting statistical tests instead of comparing just average CCRs. The statistical tests are more accurate than the average CCRs, because they employ the variance of CCR. The accuracy of the statistical tests depends on the accuracy of the estimate of the variance of CCR during cross-validation repetitions. In this context, an estimate of the variance of CCR, which is more accurate than the sample dispersion was proposed.

Initially, a theoretical analysis is undertaken assuming that the number of correctly classified utterances by any classifier in a cross-validation repetition is a realization of a hypergeometric random variable. An estimate of the variance of an hypergeometric r.v. is used to yield an accurate estimate of the variance of the number of correctly classified utterances. Although, our research was focused on cross-validation, a similar analysis can be conducted for bootstrap estimates of the CCR as well. The proposal to use the hypergeometric distribution instead of the binomial one can be considered as an extension of the work in [21]. In Dietterich's work, it is mentioned that the binomial model does not measure variation due to the choice of the training set. The

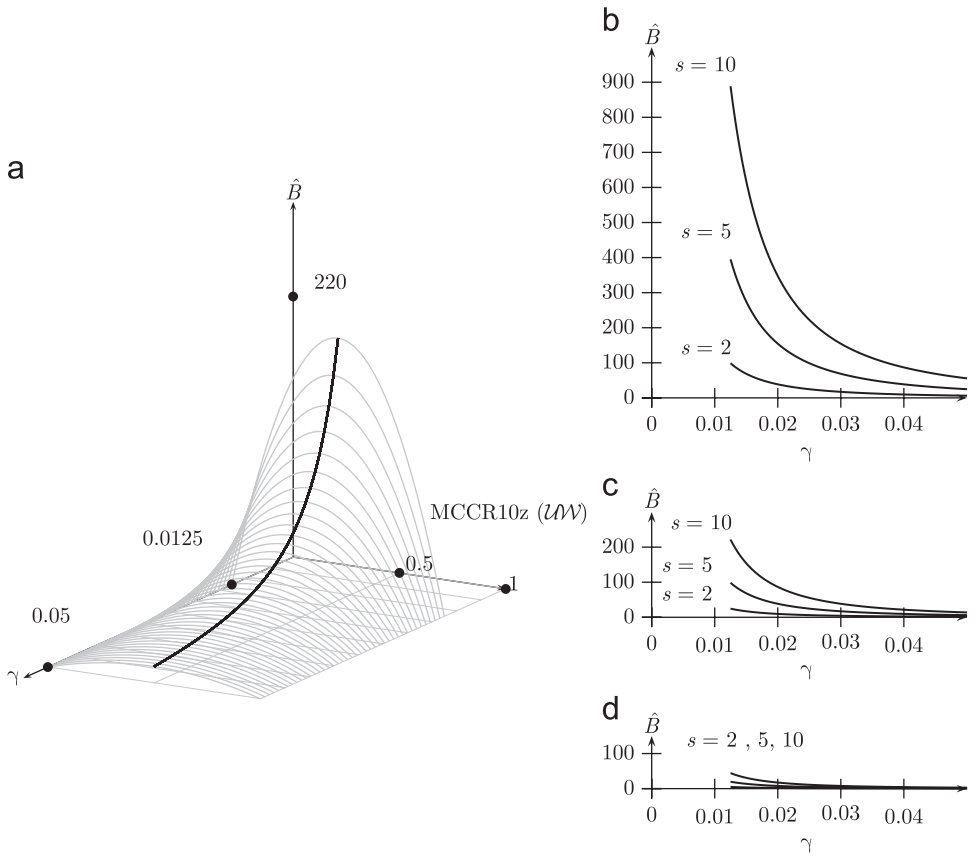

Fig. 13. $\hat{B}$ given by (44) as a function of $\mathrm{MCCR}_{10}(\mathscr{U}^{\mathscr{W}})$, γ, s, and N. In (a) free variables are γ and $\mathrm{MCCR}_{10}(\mathscr{U}^{\mathscr{W}})$. The maximum B is observed for $\mathrm{MCCR}_{10}(\mathscr{U}^{\mathscr{W}}) = 0.5$, which is plotted with a black line. For $\mathrm{MCCR}_{10}(\mathscr{U}^{\mathscr{W}}) = 0.5$, $\hat{B}$ is plotted for various s, N, and γ values in (b)–(d). (a) $s = 10$ and $N = 1000$, (b) $N = 250$, (c) $N = 1000$ and (d) $N = 5000$.

hypergeometric distribution adopted in our work remedies this variation, when the training and test sets are chosen with cross-validation.

Next, the number of correctly classified utterances committed by the Bayes classifier, when each class conditional pdf is distributed as a multivariate Gaussian, is modeled by the aforementioned hypergeometric r.v. An estimate of the variance of the CCR was derived by using the fact the CCR and the number of correctly classified utterances are strongly connected. Obviously, the variance of the CCR is limited neither by the choice of the classifier nor the pdf modeling.

Finally, the speed and the accuracy of SFFS was optimized by two methods. Method A improves the speed of SFFS with a preliminary test to avoid too many cross-validation repetitions for features that potentially do not improve the CCR. Method B improves the accuracy of SFFS by predicting the number of cross-validation repetitions, so that the confidence intervals of the CCR estimate are set to a user-defined constant. Method B controls the number of cross-validation repetitions so as the estimate of correct classification rate and its confidence limits vary less than the standard SFFS. The improved accuracy of the proposed method B is also a result of the novel estimate of the variance for the hypergeometric r.v. which varies many times less than the sample dispersion. An issue for further research is the comparison of various feature selection strategies, such as backward or random selection, with respect to the improved confidence intervals found with the proposed method. Obviously, the proposed technique is not limited to 90 features, but could handle as many features one wishes to extract from the utterances. To validate the theoretical results, experiments have been conducted for speech classification into emotional states as well as for artificially generated samples. First, it is shown that the proposed method finds an estimate that varies many times less than the sample dispersion. Second, in order to demonstrate the improvement in speed and accuracy of SFFS by the proposed methods A and B, the selection of prosody features for speech classification into emotional states was elaborated. It is found that the proposed method A reduces the executional time of SFFS by 50% without deteriorating its performance. Method B improves the accuracy of SFFS by exploiting confidence intervals of MCCR for the comparison of features. Accurate CCR values in SFFS enables the study of the 'curse of dimensionality' effect. A topic that could be further investigated.

Acknowledgment

This work has been supported by project 01ED312 co-funded by the European Union and the Greek Secretariat of Research and Technology (Greek Ministry of Development) of the Operational Program for Competitiveness within the 3rd Community Support Framework.

References

[1] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, J.G. Taylor, Emotion recognition in human–computer interaction, IEEE Signal Process. Mag. 18 (1) (2001) 32–80.
[2] M. Pantic, L.J.M. Rothkrantz, Toward an affect-sensitive multimodal human–computer interaction, Proc. IEEE 91 (9) (2003) 1370–1390.
[3] P. Oudeyer, The production and recognition of emotions in speech: features and algorithms, Internat. J. Human–Computer Studies 59 (2003) 157–183.
[4] K.R. Scherer, Vocal communication of emotion: a review of research paradigms, Speech Communication 40 (2003) 227–256.
[5] P.N. Juslin, P. Laukka, Communication of emotions in vocal expression and music performance: different channels, same code?, Psychol. Bull. 129 (5) (2003) 770–814.
[6] D. Ververidis, C. Kotropoulos, Emotional speech recognition: resources, features, and methods, Speech Communication 48 (9) (2006) 1162–1181.
[7] R. Kohavi, G. John, Wrappers for feature subset selection, Artificial Intell. 97 (1997) 273–324.
[8] A. Jain, D. Zonger, Feature selection: evaluation, application, and small sample performance, IEEE Trans. Pattern Anal. Machine Intell. 19 (2) (1997) 153–158.
[9] H. Liu, L. Yu, Toward integrating feature selection algorithms for classification and clustering, IEEE Trans. Knowledge Data Eng. 17 (2005) 491–502.
[10] F.J. Ferri, P. Pudil, M. Hatef, J. Kittler, Comparative study of techniques for large scale feature selection, in: J.E. Moody, S.J. Hanson, R.L. Lippmann (Eds.), Pattern Recognition in Practice IV, Elsevier, Amsterdam, 1994, pp. 403–413.
[11] P. Pudil, J. Novovicova, J. Kittler, Floating search methods in feature selection, Pattern Recognition Lett. 15 (1994) 1119–1125.
[12] M. Stone, Cross-validatory choice and assessment of statistical predictions, J. Roy. Stat. Soc. (Series B) 36 (2) (1974) 111–147.
[13] P. Burman, A comparative study of ordinary cross-validation, *v*-fold cross-validation and the repeated learning-testing methods, Biometrika 76 (3) (1989) 503–514.
[14] M. Evans, N. Hastings, J. Peakock, Statistical Distributions, third ed., Wiley, New York, 2000.
[15] D. Ververidis, C. Kotropoulos, Fast sequential floating forward selection applied to emotional speech features estimated on DES and SUSAS data collections, in: Proceedings of the European Signal Processing Conference (EUSIPCO), 2006.
[16] W. Nicholson, On the normal approximation to the hypergeometric distribution, Ann. Math. Stat. 27 (2) (1956) 471–483.
[17] A. Papoulis, S.U. Pillai, Probability, Random Variables, and Stochastic Processes, fourth ed., McGraw-Hill, New York, 2002.
[18] K. Fukunaga, Introduction to Statistical Pattern Recognition, second ed., Academic Press, New York, 1990.
[19] I.S. Engberg, A.V. Hansen, Documentation of the Danish emotional speech (DES) database, Internal AAU report, Center for Person Kommunikation, Aalborg University, Denmark, 1996.
[20] J.H.L. Hansen, Analysis and compensation of speech under stress and noise for environmental robustness in speech recognition, Speech Communication 20 (1996) 151–173.
[21] T. Dietterich, Approximate statistical tests for comparing supervised classification learning algorithms, Neural Computation 10 (1998) 1895–1923.