A METHOD OF AUTOMATIC SPEAKER RECOGNITION IN OPEN SETS

WOJCIECH MAJEWSKI, CZESŁAW BASZTURA, JERZY JURKIEWICZ

Institute of Telecommunication and Acoustics
Technical University of Wrocław, Poland

ABSTRACT

The paper presents a method of automatic speaker recognition in open sets ensuring a good effectiveness of elimination of strangers' voices, i.e. the voices that do not belong to a given set of known speakers. The applied procedure is discussed and description of speaker recognition experiments based on this procedure presented. The results obtained for a test material consisting of speech samples produced by 10 known speakers and 10 other speakers are very promissing /99 % of correct elimination of strangers' voices/ and confirming the pertinence of theoretical assumptions.

INTRODUCTION

In tasks of automatic speaker recognition such situations may occur in which it cannot be assumed that an unknown voice to be recognized belongs to a known set of classes of voices /closed set/. Thus, a problem arises to work out an algorithm of recognition that could operate in open sets of speakers, i.e. with no assumption that a speech sample of an unknown speaker must belong to one speaker from a given set of speakers. The idea of such approach to the problem of automatic speaker recognition was presented to the 10 ICPhS [1]. The present paper contains the analysis of this problem taking as a basis the classical Bayes's decision criterion. One of the main purposes of this study was to perform the analysis of probability of error and risk connected with a decision-making process in open sets with regard to the selection of discrimination threshold and the manner of approximation of conditional distribution of strangers' voices.

THEORETICAL BASES

In automatic voice recognition speech samples are represented by their patterns, i.e. multidimensional vectors of parameters in observation space X^K /K - space dimension/. The vectors x extracted from speech samples of particular speakers form distributions characterized by densities of conditional probabilities $Q(x|m)$, where m is a speaker number or generally a class. It may be assumed that these distributions are normal distributions expressed by the formula /Fig.1/:

$$Q(x|m)=(2\pi)^{-\frac{K}{2}}|B_m|^{-\frac{1}{2}}\exp\left\{-\frac{1}{2}(x-W_m)^{Tr}B_m^{-1}(x-W_m)\right\} \quad /1/$$

where B_m - covariance matrix for a class

W_m - mean vector for a class

$$W_m = \frac{1}{M}\sum_{i=1}^{I_m} x_{m,i}$$

$m=1,2,..M$ M - number of classes

$i=1,2,..I_m$ I_m - number of utterance repetitions for a class

Tr - sign of vector transposition

In recognition process the classical Bayes's decision criterion considers a probability $P(m|y)$ with which a test pattern y represents the class m.

$$P(m|y) = \frac{Q(y|m)P_m}{\sum_{l=1}^{M} Q(y|l)P_l} \quad /2/$$

where P_m - probability of appearances of patterns from a given class

The classical approach to recognition problem relies on finding a minimal risk $R_m(y)$ connected with assigning the pattern y to the class m.

$$R_m(y) = \sum_{l=1}^{M} C_{m,l} Q(y|l) P_l \qquad /3/$$

where $C_{m,l}$ - element of decision matrix representing the cost of decision resulting from assigning the pattern from the class l as belonging to the class m [2].

In case of speaker recognition in open sets the set of classes consists of M known classes /closed set/ and one multiobject class corresponding to all other voices that do not belong to the set M. These voices constitute so called "ground" or strangers' voices class /m = 0/. The conditional distribution $Q(x|0)$ of ground is in general case a multimodal distribution with parameters that are not known.

Considering these assumptions the recognition procedure in open sets may be presented as consisting of two stages:

1. Identification in the closed set, i.e. finding m^* for which

$$R_{m^*}(y) = \min_{m} R_m(y) \qquad /4/$$

what means a temporary assigning a test pattern y to the class m^*.

2. Verification, i.e. checking the condition

$$R_{m^*}(y) < R_0(y) \qquad /5/$$

If the condition /5/ is fulfilled, the pattern y belongs to the class m^* ; in the opposite case it belongs to the class m = 0, i.e. the ground.

It is to observe that the formula /4/ permits to devide the parameter space into M subspaces X_m^K. Similarly, the inequality /5/ defines areas X_{Wm}^K in subspaces X_m^K /Fig.2/. It follows that it is not necessary to know the total distribution of $Q(x|0)$, but only the limits of areas X_{Wm}^K or $Q(x|0)$ distribution in the vicinity of these limits. Thus, the $Q(x|0)$ distribution may be approximated by means of M planes, one by one for each subspace X_m^K.

$$Q(x|0) = G_m(x) \qquad m : \; x \in X_m^K \qquad /6/$$

where $G_m(x) = g_{m,0} + \sum_{k=1}^{K} g_{m,k} + x_k \qquad /7/$

is the equation of plane m in subspace X_m^K /Fig.2/. Discontinuities of such approximation at the borders of subspaces X_m^K are insignificant for the verification process.

RECOGNITION ERRORS

The information about errors is contained in the statistics of identification and verification shown in Fig.3. In this figure particular symbols have the meaning:

N - number of voice patterns
W - patterns belonging to the closed set
O - patterns from beyond the closed set
P - initially correctly recognized
B - initially incorrectly recognized
A - accepted by verification
E - eliminated by verification

For example N_{WPE} indicates the number of patterns from the closed set, correctly recognized by the classifier, but next rejected in the verification process.

Within the closed set the statistics of incorrect recognitions is represented by:

$$\delta = \frac{N_{WB}}{N_W} \qquad /8/$$

Verification procedure devides this error into two components:

$$\delta_A = \frac{N_{WBA}}{N_W} \quad \text{and} \quad \delta_E = \frac{N_{WBE}}{N_W} \qquad /9/$$

and introduces verification errors: the error of incorrect rejection expressed as

$$\alpha' = \frac{N_{WPE}}{N_W} \qquad /10/$$

or as $$\alpha'' = \frac{N_{WPE} + N_{WB}}{N_W} = \alpha' + \delta \qquad /11/$$

$$\alpha''' = \frac{N_{WPE} + N_{WBE}}{N_W} = \alpha' + \delta_E \qquad /12/$$

and - in relation to the open set - also the error of false acceptance:

$$\beta = \frac{N_{OA}}{N_O} \qquad /13/$$

EXPERIMENTAL PROCEDURE

Speaker recognition experiments in open set were performed in the following conditions:

a/ A specific cue material was used. It was a Polish sentence "Jutro będzie ładny dzień" /Tomorrow it'll be a fine day/. Distributions of time intervals between zero-crossings [3] were extracted from this sentence and used as vectors of parameters. The dimension of observation space was K = 4 /the parameters of the largest discrimination power were selected/.

b/ The learning sequence consisted of 100 vectors $x_{m,i}$ obtained from I_m = 10 repetitions of the utterance by M = 10 speakers.

c/ The testing sequence consisted of 10 other repetitions of the utterance by 10 speakers from the closed set and 10 repetitions by 10 speakers from beyond the closed set. Thus, the open set contained 200 vectors $y_{m,i}$ obtained from 20 speakers.

Since the main concern of this study was verification procedure for a fixed measurement set-up, the experiments were arranged in such a way that first speaker identification procedure was applied to the total testing sequence and next verification procedure was utilized for different values of verification parameters.

In the recognition process $m^*_{m,i}$, $R_{m^*}(y_{m,i})$ and $R_0(y_{m,i})$ were calculated for each pair m,i /see formulas 3,4 and 5/, assuming that the elements $C_{m,l}$ of matrix C are equal 1 in case of incorrect decisions or 0 in case of correct decisions.

From the possible ways of $Q(x|0)$ approximation /eq. 6 and 7/ two simple cases were investigated in the experiments /see Fig.2/:

1^o $Q(x|0) = H \quad H = \text{const} \qquad /14/$

2^o $Q(x|0) = H_m \quad m : x \in X_m^K \qquad /15/$

For the first case the decision threshold was defined as

$$H = \gamma Q_{av} \qquad Q_{av} = \frac{1}{M} \sum_{m=1}^{M} Q(W_m|m) \qquad /16/$$

where γ - coefficient /experiment parameter/

For the second case two versions of defining individual thresholds H_m were distinguished:

$$H_m = \gamma Q(W_m|m) \qquad /17/$$

and $$H_m = \gamma_m Q(W_m|m) \qquad /18/$$

where γ_m - coefficient selected individually to minimize verification risk for a given class on the basis of α' and β errors.

RESULTS AND CONCLUSIONS

The results of the experiments are set together in Table 1 which presents the errors for different approximations of $Q(x|0)$ and γ values that minimize the verification risk. The influence of γ coefficient on α' and β errors for the case nr 1 /eq. 14 and 16/ is shown as example in Fig. 4. Analyzing the data presented in Table 1 it may be noticed that the speaker recognition scores are very little differentiated in the examined cases. This may be the result of very effective discriminating power of the vectors applied and/or too small size of the test set. It is, however, necessary to emphasize that the results obtained confirmed the pertinence of methodological assumptions what was the main purpose of this study.

The methodological considerations permit to state that the proposed method of voice recognition in open sets is very elastic and it enables to adjust the global characteristics, i.e. α and β errors, to adopted strategy of recognition system. For a given set of patterns describing the voices it is always possible to optimize the recognition by a proper selection of approximation of the ground class distribution, i.e. by proper selection of decision threshold. It is a basic advantage of the presented method of speaker recognition in open sets verified experimentally for a test population of 20 speakers.

Table 1. Recognition errors

Case	1°-eq16 $\gamma=2.10^{-5}$	2°-eq17 $\gamma=2.10^{-4}$	2°-eq17 $\gamma=3.10^{-4}$	2°-eq18 ind γ_m
Error	%	%	%	%
δ	8	8	8	8
δ_A	6	6	6	6
δ_E	2	2	2	2
α^I	1	2	1	1
α^{II}	9	10	9	9
α^{III}	3	4	3	3
β	1	1	2	1

REFERENCES

1. W.Majewski, Cz.Basztura, Speaker recognition in open sets, Proceedings of the Tenth International Congress of Phonetic Sciences /M.P.R. Van den Broecke and A.Cohen eds./, Foris Publications, Dordrecht, 1984, 322-325.

2. J.Z.Cypkin, Podstawy teorii układów uczacych sie, WN-T, Warszawa, 1973.

3. Cz.Basztura, W.Majewski, The application of long-term analysis of the zero-crossing of a speech signal in automatic speaker identification, Archives of Acoustics, 3, 1, 1978, 3-15.

FIGURES

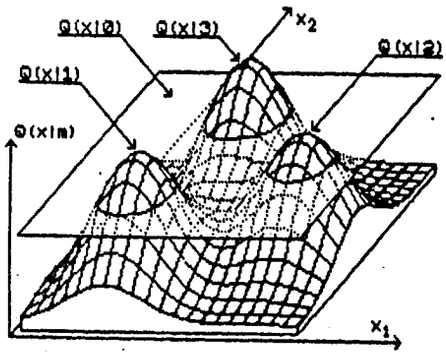

Fig.1. Examples of $Q(x|m)$ distributions in case of two dimensional space /K = 2/.

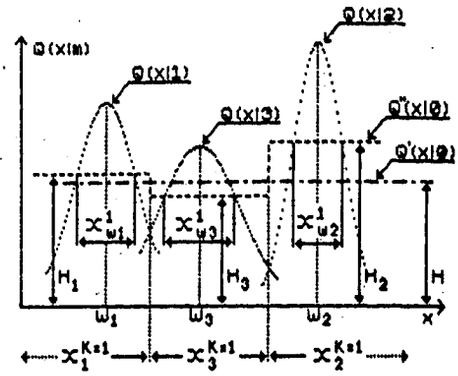

Fig.2. Illustration of approximation of $Q(x|0)$ and determination of decision threshold H for one dimensional space.
$Q'(x|0)$ - case nr 1 /eq.14/
$Q''(x|0)$ - case nr 2 /eq.15/

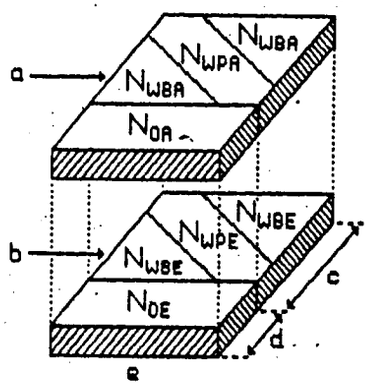

Fig.3. Statistics of recognitions in open sets; a - recognitions accepted by verification, b - recognitions rejected by verification, c - patterns from the closed set, d - patterns from beyond the closed set, e - classes.

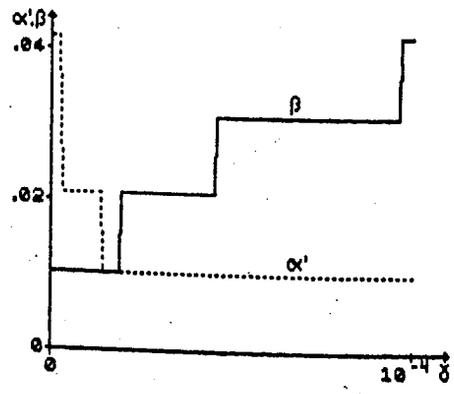

Fig.4. α and β errors in the function of γ for the case nr 1 /eq.14 and 16/.