Speaker Recognition in Open Sets

W. Majewski and C. Basztura
Wrocław, Poland

1. Introduction

The development of methods and techniques for automatic speaker recognition creates a real promise for their future practical applications in a variety of social and business activities. These include such fields as criminology, protection of information stored in computer memory, voice control of machines by authorized persons. In all these applications a problem of speaker verification or identification may occur. The term verification refers to a decision-making process leading to an acceptance or rejection of claimed identity of a speaker (Rosenberg, 1976). Similarly, the term identification – in most common understanding (Rosenberg, 1976) – refers to a process that assigns a speech sample of an unknown speaker to one speaker from a given set of speakers. This classical approach to speaker identification makes sense only in case of closed sets of speakers, i.e. such sets in which it is a priori assumed that the unknown speaker belongs to the given set. Such presumption is, however, an oversimplification of the speaker identification problem. If the set of speakers is open, the unknown speaker may not necessarily belong to the set of known speakers, and the identification task has to be preceded by a corroboration that the unknown speaker does in fact belong to the set of known speakers. Thus, the question arises how the recognition system should react and what decision criteria should be adopted to find out whether the speech pattern of an unknown speaker belongs to the given set of M speakers or is a representation of a new, M + 1 class of voices. The present study, which is part of a substantial series of investigations focused on speaker identification (e.g. Majewski et al. 1981; Basztura and Majewski, 1981), is an attempt to answer these important questions.

2. Recognition System — Preliminary Assumptions

A model of a recognition system consisting of two basic operational blocks, for measurement and for classification, was adopted for this study. It is a simple system with programmed learning and statistical decision criteria (Skripkin and Gorielik, 1977) Reference patterns W_m are developed by means of learning sequences representing I_m repetitions of speech samples produced by each speaker belonging to the set of M classes. The assumption

of a system with programmed learning fits the majority of known, practically applied recognition systems (Rosenberg, 1976; Atal, 1976) and decision criteria based on statistics (Skripkin and Gorielik, 1977).

The main goal of speaker recognition procedure is to assure a minimum of incorrect decisions, which may be classified as α and β errors. The first kind of error is called a *false alarm* and it defines the chance of incorrect rejection. β errors, or *misses*, define the probability of incorrect identifications. It should be emphasised that the efficiency of a recognition system, i.e. minimization of both α and β errors, depends on the selection of effective parameters representing the individual voice features in suitably chosen utterances (Wolf, 1972). The expected value of correct decisions is also a function of population size M (Rosenberg, 1976) and representativeness of learning sequence. The error rates depend also on the classification stage, which is the main concern of this study. Thus, the present discussion concerns a fixed measuring procedure and is aimed at a classification procedure that permits avoiding an incidental increase of β errors in cases of open sets of test speech samples.

The classical approach to speaker identification does not assure such features of classification, since the pattern X_n of unknown speaker is assigned to the particular class m* from the set of M classes, if the adopted similarity measure between this pattern and reference patterns of all classes reaches a maximum for the class m*.

In terms of distance measures:

$$\left[\begin{array}{l} X_n \in m^* \ \text{if} \ \overset{M}{\underset{m=1}{\mathrm{Min}}}\left[d(X_n,W_m)\right] = d(X_n,W_{m^*}) \\ \text{where } m = 1,2,\ldots M \quad M - \text{number of classes} \end{array}\right.$$

It is clear that the application of the classical approach to open sets of speakers also assigns the pattern of unknown voice to some class from the set M in case where the test sample does not belong to any of the known classes.

One possible conception that permits avoiding the stated inconvenience of the classical approach seems to be the idea of reducing the problem of speaker recognition in open sets to the known and successfully solved problem of speaker verification. A speaker recognition system is presented in the next section of the paper based on this idea.

3. Proposed Model of the Recogniton System

The proposed system is based on a dichotomous classification (Cypkin, 1973) with the utilization of statistical decision theory, in one case followed by the verification task.

The operation of the system, which enables us to reduce the problem of speaker recognition in open sets to the problem of speaker verification, may be presented in the following points.

A. Learning cycle

- For a number of speakers M, and learning sequenc I_m, establish the kind of parameters extracted from the speech signal, and resulting from them patterns X of individual voice features, the probability P_m of appearances of patterns from a given class, and adopt a particular similarity measure between speech patterns as well as preliminary assumptions related to expected values of α and β errors, which should be facilitated by the introductory experiments.
- Derive estimators of the probability distributions $p_m(X)$ for each of all the known M classes of voices.
- On the basis of assumed loss functions c and probabilities P_m derive threshold values h for each pair of classes from the set M.

$$\left[h_{m,m+1} = \frac{c_{m+1,m} - c_{m+1,m+1}}{c_{m,m+1} - c_{m,m}} \; \frac{P_{m+1}}{P_m} \right.$$

- For each class from the set M derive the reference pattern W_m and decision threshold H_m as in the case of a simple discrimination task.

B. Classification cycle

- Read in the parameters of a pattern X_n from an unknown speaker.
- For an arbitrary chosen pair of classes compute the likelihood ratio

$$\left[l_{m,m+1}\left(X_n\right) = \frac{p_m \; X_n}{p_{m+1} \; X_n} \right.$$

  and compare it with the corresponding threshold value derived in the learning cycle. If $l_{m,m+1}$ (Xn) > $h_{m,m+1}$, the pattern X_n from the unknown speaker is 'closer' to the class m than to class m + 1, which is rejected. In the opposite case the pattern X_n is closer to the class m + 1 and class m is rejected.
- Replace the rejected class with the values $P_{m+2}(X_n)$ from the next class and compute a subsequent likelihood ratio and compare it with the corresponding threshold value. In the course of checking all known M classes the index m* of the class closest to the pattern X_n is established.
- Carry out the verification procedure, i.e. compare the value of the similarity measure between the test pattern X_n and reference pattern W_{m*} with the decision threshold H_m* derived in the learning cycle. Thus:

$$\left| \begin{array}{l} X_n \in m^* \;\; \text{if} \;\; d\left(X_n, W_m *\right) < H_m * \\ X_n \notin m^* \;\; \text{if} \;\; d\left(X_n, W_m *\right) \geqslant H_m * \end{array} \right.$$

In the first case, the pattern X_n with the assumed probability of error is assigned to the class m*, in the second – the pattern X_n does not belong to any class from the set M and may be stored as a pattern generating a new, M + 1 class of voices.

4. Concluding Remarks

In comparison to the classical approach to speaker identification the presented procedure is more complicated and requires more computations. However, considering the actual state of the art in computer hardware and software, this should not be a serious obstacle in practical applications of this procedure. On the other hand, the proposed system is more universal and it enables realization of a speaker identification procedure both in closed and open sets of speakers, as well as realization of a speaker verification procedure. The application of the present system permits performing the speaker recognition task with an established, for any given input data, probability of correct decision, which is impossible to achieve when a classical decision rule of speaker identification is applied to open sets of speakers.

References

Atal, B.S. (1976). Automatic recognition of speakers from their voices. *Proceedings of the IEEE,* **64**, 460-475.

Basztura, C. and Majewski W. (1981). The effect of chosen parameters of a telephone channel on voice identification. *Archives of Acoustics,* **6**, 359-370.

Cypkin, J.Z. (1973). *Podstawy teorii ukladów uczacych sie.* Warszawa: Wydawnictwa Naukowo-Techniczne.

Majewski, W., Tosi, O. and Jurkiewicz, J. (1981). Speaker identification over telephone lines with unknown frequency response. *Proceedings of the Fourth FASE Symposium on Acoustics and Speech, Venezia,* **I**, 313-316.

Rosenberg A.E. (1976). Automatic speaker verification: a revue. *Proceedings of the IEEE,* **64**, 475-487.

Skripkin, W.A. and Gorielik, A.L. (1977). *Mietody rozpoznavanija.* Moskva: Visshaia Shkola.

Wolf, J.J. (1972). Efficient acoustic parameters for speaker recognition. *Journal of the Acoustical Society of America,* **51**, 2044-2056.