AN APPROACH TO SPEAKER IDENTIFICATION USING MULTIPLE CLASSIFIERS

Vlasta Radová and Josef Psutka

University of West Bohemia, Department of Cybernetics
Univerzitní 22, 306 14 Plzeň, Czech Republic
radova@kky.zcu.cz, psutka@kky.zcu.cz

ABSTRACT

Presented paper takes interest in a speaker identification problem. The attributes representing voice of a particular speaker are obtained from very short segments of the speech waveform corresponding only to one pitch period of vowels. The patterns formed from the samples of a pitch period waveform are either matched in time domain by use of a nonlinear time warping method, known as dynamic time warping (DTW), or they are converted into the cepstral coefficients and compared using the cepstral distance measure. Since an uttered speech signal usually contains a lot of vowels the techniques using a combination both various classifiers and multiple classifier outputs are considered in the decision making process. Experiments performed for hundred speakers are described at the end of this paper.

1. INTRODUCTION

Speaker recognition methods can be divided into three groups with the view to the presumption where in the pronounced speech signal the information about the individuality of a speaker is encoded and in what kind of manner. The basic division can be presented as

- recognition using time functions of suitable speech parameters,
- recognition on the basis of the long-term averages of suitable speech parameters,
- recognition with the searching for specific phonetic events.

The speaker recognition approach based on using time functions of suitable parameter is mostly used in the text-dependent systems. The information about individuality of a speaker is supposed to be contained in the manner of the pronunciation of a formerly chosen utterance. The pronounced utterance is first processed by a speech processing method and then represented by the time function of the chosen parameter. The recognition of an unknown speaker is based on matching the time function of the unknown speaker with the time function of the reference speaker [1], [2].

The techniques using long-term averages of the suitable speech parameters are typically used in the text-independent speaker recognition systems. The function of such systems is based on the hypothesis that the information about a speaker is present in a component of the speech signal that has for the particular speaker fixed average value given by the anatomy of his vocal tract. The divergences are expected to be brought about only by the phonetical variability of particular pronounced utterances. As the result of these considerations it is supposed that the information about the speaker can be obtained from any speech signal by averaging a chosen parameter. The process of recognition is then performed by matching the average parameter values obtained from the utterance of unknown speaker with the stored long-term averages of reference speakers [2], [3].

Speaker recognition based on searching for specific phonetic events can be used both in text-dependent and text-independent tasks. The fundamental idea of this approach is to find in the speech signal such phonetic events that are specific for the given speaker. Such events can be for example vowels [4], nasals or signal segments corresponding to transition coarticulatory effects arising between nasals and vocals [5] etc. The recognition is based on a comparison the patterns extracted from the specific phonetic events of the unknown speaker with patterns belonging to the phonetic events of the reference speaker.

Our paper concerns just in a problem of searching for specific phonetic events. Since a speech signal can contain a lot of such events suitable decision making techniques are investigated.

The patterns representing specific phonetic events are formed from the very short parts of the vowels the length of which corresponds only to one pitch period of the speech waveform [6]. It means that directly the raw speech samples are regarded as features. Every speaker (both reference and unknown) is represented by a set of such patterns (current utterance contains usually several vowels) and every pattern is classified by two different classifiers.

One classifier compares the patterns using the nonlinear time warping method (DTW). The other classifier first converts the patterns into vectors of cepstral coefficients and then classifies these vectors using the cepstral measure. The final decision about the identity of the unknown speaker is determined by a combination of outputs of these two classifiers obtained for all vowels that were taken into account for given utterance.

2. DESCRIPTION OF THE IDENTIFICATION ALGORITHM

Suppose that the unknown speaker, represented by a set of patterns $X = \{ \boldsymbol{x}_{hl} \mid l = 1, 2, \ldots, L_h,\ h = 1, 2, \ldots, H \}$, where $\boldsymbol{x}_{hl}$ is the l-th pattern of the unknown speaker obtained from the vowel h, L_h is the number of patterns obtained from the vowel h and H is the number of various vowels from which the patterns of the unknown speaker were obtained, should be identified as one of M reference speakers. Further suppose that there are K partial classifiers φ_k, $k = 1, 2, \ldots, K$, (in our experiments K=2) each of them assignes each particular pattern $\boldsymbol{x}_{hl} \in X$ either one index $i_{khl} \in \Lambda$, $\Lambda = \{1, 2, \ldots, M\}$, as a label that $\boldsymbol{x}_{hl}$ belongs to the class $\omega_{i_{khl}}$ or the index $i_{khl} = M+1$ if the classifier φ_k is not able to decide which class the pattern $\boldsymbol{x}_{hl}$ belongs to. Such classifiers provide so-called information of abstract level [7] and their function may be described as

$$\varphi_k(\boldsymbol{x}_{hl}) = i_{khl} \ , \tag{1}$$

where $i_{khl} \in \Lambda \cup \{M+1\}$, $\Lambda = \{1, 2, \ldots, M\}$ and M is the number of reference speakers. Since conflicts may exist among the decisions of the partial classifiers achieved for the particular patterns $\boldsymbol{x}_{hl}$, $l = 1, 2, \ldots, L_h$, $h = 1, 2, \ldots, H$, it is necessary to design a general classifier Φ that will use all i_{khl}, $k = 1, 2, \ldots, K$, $l = 1, 2, \ldots, L_h$, $h = 1, 2, \ldots, H$, from (1) to recognize the unknown speaker (represented by the set X) as one of the M reference speakers, i.e.

$$\Phi(X) = i \ , \tag{2}$$

where $i \in \Lambda \cup \{M+1\}$, $\Lambda = \{1, 2, \ldots, M\}$ and M is the number of reference speakers. A simple and common rule used for resolving this kind of conflicts in human social life is voting by majority. This rule can be expressed by the formula

$$\Phi(X) = \begin{cases} i^*, & \text{if } \exists!\, i^* \in \Lambda : N(i^* \mid X) = \max\limits_{i \in \Lambda} N(i \mid X), \\ M+1 & \text{otherwise} \end{cases} \tag{3}$$

where

$$N(i \mid X) = \sum_{k=1}^{K} \sum_{h=1}^{H} \sum_{l=1}^{L_h} \tau_k(\boldsymbol{x}_{hl} \in \omega_i) \tag{4}$$

is the number of votes for the class ω_i, $i \in \Lambda$, and

$$\tau_k(\boldsymbol{x}_{hl} \in \omega_i) = \begin{cases} 1, & \text{if } \varphi_k(\boldsymbol{x}_{hl}) = i \text{ and } i \in \Lambda \ , \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

is a binary characteristic function representing the event $\varphi_k(\boldsymbol{x}_{hl}) = i$.

Another rules for the combination of multiple decisions may be used if we suppose that each partial classifier φ_k, $k = 1, 2, \ldots, K$, provides so-called information of rank level [7], [8]. In this case the partial classifier φ_k may be described as

$$\varphi_k(\boldsymbol{x}_{hl}) = \Omega_{khl} \tag{6}$$

where Ω_{khl} contains all labels $i \in \Lambda$ ranked in a queue with the label at the top being the best choice. The combination of outputs of such classifiers may be based on either Condorcet consistent rules or scoring methods [9]. A typical representative of Condorcet consistent rules is the Condorcet winner rule

$$\Phi(X) = \begin{cases} i^*, & \text{if } \exists\, i^* \in \Lambda : N(i^* \prec j \mid X) > N(j \prec i^* \mid X) \\ & \qquad \forall\, j \in \Lambda - \{i^*\} \ , \\ M+1 & \text{otherwise} \end{cases} \tag{7}$$

where

$$N(i \prec j \mid X) = \sum_{k=1}^{K} \sum_{h=1}^{H} \sum_{l=1}^{L_h} \tau_k(i \prec j \mid \boldsymbol{x}_{hl}) \tag{8}$$

is the number of patterns $\boldsymbol{x}_{hl} \in X$ and classifiers φ_k, $k = 1, 2, \ldots, K$, which prefer class ω_i to class ω_j, $i \neq j$, $i, j \in \Lambda$, and

$$\tau_k(i \prec j \mid \boldsymbol{x}_{hl}) = \begin{cases} 1, & \text{if } z_k(i \mid \boldsymbol{x}_{hl}) < z_k(j \mid \boldsymbol{x}_{hl}) \ , \\ 0 & \text{otherwise} \end{cases} \tag{9}$$

is a binary characteristic function representing the event that the classifier φ_k prefers class ω_i to class ω_j for the pattern $\boldsymbol{x}_{hl}$ and $z_k(i \mid \boldsymbol{x}_{hl})$ is a function representing the rank of the class ω_i, $i \in \Lambda$, in the Ω_{khl}. A typical representative of scoring methods is the Borda rule

$$\Phi(X) = \begin{cases} i^*, & \text{if } \exists!\, i^* \in \Lambda : i^* = \operatorname{argmax}\limits_{i \in \Lambda} S(i \mid X) \ , \\ M+1 & \text{otherwise} \end{cases} \tag{10}$$

where

$$S(i \mid X) = \sum_{k=1}^{K} \sum_{h=1}^{H} \sum_{l=1}^{L_h} S_{\mathrm{Bk}}(i \mid \boldsymbol{x}_{hl}) \tag{11}$$

is the total score of the class ω_i, $i \in \Lambda$, and

$$S_{\mathrm{Bk}}(i \mid \boldsymbol{x}_{hl}) = M - z_k(i \mid \boldsymbol{x}_{hl}) \tag{12}$$

is so-called Borda score.

3. DESCRIPTION OF THE CLASSIFIERS

Both classifiers mentioned in Section 1 operate with the patterns composed of samples of one pitch period of speech waveform. To avoid the differences in the amplitude all patterns are normalized in such a way that

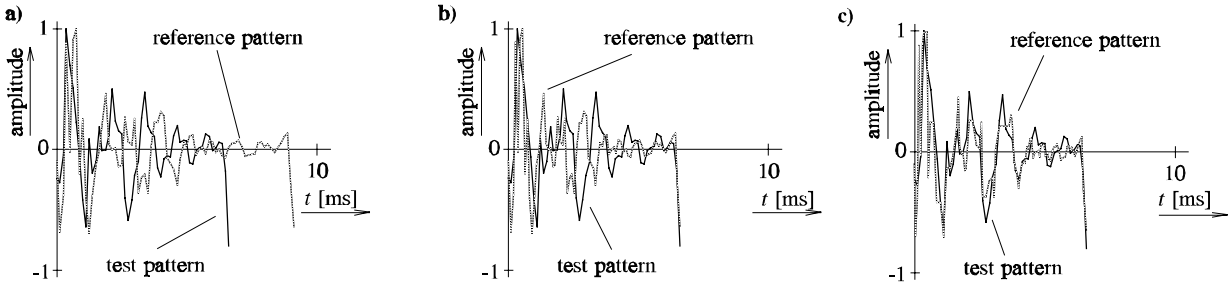

Fig. 1. An illustration of the patterns alignment. a) The original patterns, b) the patterns after linear warping that aligns endpoints, c) the patterns after nonlinear alignment using the DTW.

the absolute value of the maximum amplitude of each normalized pattern is equal to 1. The classification of the patterns is based on the nearest neighbour principle, i.e. the pattern $\boldsymbol{x}_{hl}$ is classified into the class ω_{i^*}, $i^* \in \Lambda$, $\Lambda = \{1, 2, ..., M\}$, for which

$$i^* = \underset{i \in \Lambda}{\operatorname{argmin}}\ d(i, \boldsymbol{x}_{hl}) \qquad (13)$$

where $d(i, \boldsymbol{x}_{hl})$ is a distance measure between the pattern $\boldsymbol{x}_{hl}$ and the class ω_i. The distance measure, however, is defined in different way for each partial classifier.

The first classifier uses the nonlinear time warping technique, that enables to align the patterns much better than can be attained by a linear time alignment (see Fig. 1). The distance measure $d(i, \boldsymbol{x}_{hl})$ is then determined as

$$d(i, \boldsymbol{x}_{hl}) \equiv d(\boldsymbol{r}^{\,i}_{hv\cdot}, \boldsymbol{x}_{hl}) = \min_{v} D(\boldsymbol{x}_{hl}, \boldsymbol{r}^{\,i}_{hv}) \qquad (14)$$

where $\boldsymbol{r}^{\,i}_{hv}$, $v = 1, 2, ..., V^i_h$, $h = 1, 2, ..., H$, $i = 1, 2, ..., M$, is the v-th pattern of the i-th reference speaker obtained from the vowel h, V^i_h is the number of patterns of the i-th reference speaker obtained from the vowel h, H is the number of various vowels from which the patterns of the unknown speaker were obtained, M is the number of reference speakers and $D(\boldsymbol{x}_{hl}, \boldsymbol{r}^{\,i}_{hv})$ is the distance between the pattern $\boldsymbol{x}_{hl}$ of the unknown speaker and the pattern $\boldsymbol{r}^{\,i}_{hv}$ of the i-th reference speaker determined as the by-product of the DTW [10]. The type of permitted transitions of the DTW function employed in the matching process is depicted in Fig. 2.

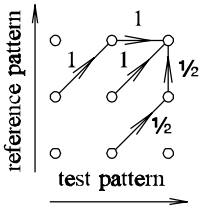

Fig. 2. Employed type of the permitted transitions of the DTW function.

The second classifier first converts each pattern into the vector of 7 LPC-derived cepstral coefficients and then classifies these vectors using the cepstral measure. The distance measure in (13) is then defined as

$$d(i, \boldsymbol{x}_{hl}) \equiv d(\boldsymbol{r}^{\,i}_{hv\cdot}, \boldsymbol{x}_{hl}) = \min_{v} \left((\boldsymbol{r}^{\,i}_{hv} - \boldsymbol{x}_{hl})^{\mathrm{T}} (\boldsymbol{r}^{\,i}_{hv} - \boldsymbol{x}_{hl}) \right) . \qquad (15)$$

4. EXPERIMENTAL RESULTS

The described speaker identification method has been tested in the same group of 100 speakers as in [6]. Each speaker was represented by 60 patterns (12 for each of 5 Czech vowels). Thirty patterns (6 for each vowel) were regarded as reference patterns (i.e. patterns of the reference speaker), and thirty others as test patterns (i.e. patterns of the speaker to be recognized). Results of the identification experiments are shown in Table I. Both the results achieved for particular vowels and the total results are presented. The mark "+" means the number of correctly recognized speakers, "−" the number of misrecognized speakers and "?" the number of cases in which the classifier is not able to identify the tested speaker definitely.

In comparison with the results achieved for the two partial classifiers independently (Tables II and III) the results presented in Table I show a considerable increase of the number of correctly recognized speakers both for particular vowels and for all vowels in total. For example using the majority voting rule the number of correctly recognized speakers increases from 72% for the classifier with the cepstral coefficients and 88% for the classifier with the DTW to 98% for the classification method proposed in this paper. Similar situation occurs also for particular vowels and the other combination rules described in Section 2.

5. CONCLUSION

In the paper a speaker identification method has been presented that uses parts of the vowel waveform as patterns representing a particular speaker. These patterns are classified by two different classifiers and the final identification of the unknown speaker is based on a combination of outputs of these two classifiers. Using this method as many as 98% speakers in a group of 100

speakers were identified correctly. A comparison of these results with results reported by other authors is rather difficult since, to our knowledge, no experiments in a group of 100 speakers (or higher) have been reported with patterns obtained only from the vowels so far. However, since Fakotakis et al. in [4] reported 90% of correctly identified speakers in a group of 15 speakers using vowels as the identification material, the proposed method may be regarded as a promising way how to achieve a high speaker identification performance.

6. REFERENCES

[1] B. S. Atal: "Automatic Speaker Recognition Based on Pitch Contours." *J. Acoust. Soc. Amer.*, vol. 52, pp. 1687−1697, 1972.

[2] S. Furui: "Comparison of Speaker Recognition Methods Using Statistical Features and Dynamic Features." *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-29, pp. 176−182, 1981.

[3] J. D. Markel, S. B. Davis: "Text-Independent Speaker Recognition from a Large Linguistically Unconstrained Time-Spaced Data Base." *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-27, pp. 74−82, 1979.

[4] N. Fakotakis, A. Tsopanoglou, G. Kokkinakis: "A Text-Independent Speaker Recognition System Based on Vowel Spotting." *Speech Communication*, vol. 12, pp. 57−68, 1993.

[5] L.-S. Su, K.-P. Li, K. S. Fu: "Identification of Speakers by Use of Nasal Coarticulation." *J. Acoust. Soc. Amer.*, vol. 56, pp. 1876−1882, 1974.

[6] V. Radová, J. Psutka: "Speaker Recognition Using Raw Speech Data as Features." In *Sprachkommunikation*, ITG-Fachbericht 139, VDE-Verlag GmbH, Berlin, pp. 73−76, 1996.

[7] L. Xu, A. Krzyżak, C. Y. Suen: "Methods of Combining Multiple Classifiers and Their Applications to Handwriting Recognition." *IEEE Trans. Syst., Man, Cybern.*, vol. 22, pp. 418−435, 1992.

[8] T. K. Ho, J. J. Hull, S. N. Srihari: "Decision Combination in Multiple Classifier Systems." *IEEE Trans. Pattern Anal. and Machine Intell.*, vol. 16, pp. 66−75, 1994.

[9] H. Moulin: *Axioms of Cooperative Decision Making.* Cambridge University Press, Cambridge·New York· Port Chester·Melbourne·Sydney 1988.

[10] D. O'Shaughnessy: *Speech Communication. Human and Machine.* Addison-Wesley Publishing Company, Reading 1987.

Table I. Number of recognized speakers in a group of 100 speakers using the general classifier.

Number of recognized speakers [%]		vowel															total		
		/a/			/e/			/i/			/o/			/u/					
		+	−	?	+	−	?	+	−	?	+	−	?	+	−	?	+	−	?
rule	majority voting (3)	42	27	31	50	20	30	63	18	19	64	12	24	39	20	41	98	2	0
	Condorcet winner (7)	39	19	42	42	11	47	57	10	33	59	7	34	32	6	62	89	4	7
	Borda scoring (10)	58	41	1	60	40	0	70	29	1	68	31	1	44	55	1	93	7	0

Table II. Number of recognized speakers in a group of 100 speakers using the classifier with the cepstral coefficients.

Number of recognized speakers [%]		vowel															total		
		/a/			/e/			/i/			/o/			/u/					
		+	−	?	+	−	?	+	−	?	+	−	?	+	−	?	+	−	?
rule	majority voting (3)	16	37	47	24	26	50	44	30	26	33	21	46	12	27	61	72	16	12
	Condorcet winner (7)	15	14	71	24	15	61	47	12	41	25	5	70	10	6	84	80	6	14
	Borda scoring (10)	32	67	1	49	47	4	64	34	2	46	52	2	29	68	3	82	18	0

Table III. Number of recognized speakers in a group of 100 speakers using the classifier with the DTW.

Number of recognized speakers [%]		vowel															total		
		/a/			/e/			/i/			/o/			/u/					
		+	−	?	+	−	?	+	−	?	+	−	?	+	−	?	+	−	?
rule	majority voting (3)	32	25	43	44	19	37	45	23	32	57	19	24	33	23	44	88	5	7
	Condorcet winner (7)	24	10	66	32	5	63	30	10	60	51	2	47	29	12	59	85	5	9
	Borda scoring (10)	49	48	3	47	52	1	49	48	3	59	39	2	37	60	3	84	15	1