COMBINING TEMPLATE MATCHING AND MULTILAYER PERCEPTRON FOR SPEAKER IDENTIFICATION

X.Y. Zhu and L.W. Cahill
Department of Electronic Engineering, La Trobe University

ABSTRACT - This paper presents a combination approach to speaker identification. In addition to a traditional template matching method, a multilayer perceptron (MLP) method was applied to further distinguish speakers' voices. In the template matching method, cepstral coefficients were selected as acoustic features, and a dynamic time warping (DTW) algorithm was used to compare the feature vectors at equivalent points in time. An unknown speaker's template was first compared with all the stored speakers' reference templates to choose a few candidates. The MLP method, in which formant parameters of vowels and diphthongs were chosen as features, was then used on these candidates to identify the the identity of the speaker. The final results showed that the combination approach was better than either the traditional method or the MLP method, used alone.

INTRODUCTION

The task of speaker identification is to label an unknown voice as one of a set of known voices. In police work, it can be applied to associating a particular person with a voice. Like other means of biometric personal identification, speaker identification is generally considered to be more reliable than artifact identification because it is based on intrinsic characteristics of the individual which are difficult, if not impossible,to mimic. Human listeners use many different "sources of knowledge" for identifying speakers; these include high-level features such as style of speaking, accent, and vocabulary. Since a lot of information is difficult to quantify or represent, automatic speaker recognizers exploit only lower-level acoustic features which reflect vocal chord and vocal tract shape.

Speaker identification is much more difficult than speaker verification, since its performance degrades with an increasing number of users (Nail,1990). Traditional speaker identification methods are template matching and hidden Markov models. The selection of features that are effective for distinguishing speakers' voices is crucial in the success of speaker recognition. Cepstral coefficients and formants of certain vowels and diphthongs have long been recognized as very effective features, and are used in a number of speaker identification algorithms (Sethuraman & Gowdy, 1989) (Miles & Guillemin, 1989). Recently, connectionist models have been applied to speech signal processing, mostly in speech recognition and feature extraction. Their main useful properties are their discriminative power and their ability to capture input-output relationships. Templeton and Guillemin presented an MLP approach to speaker identification, and showed that the MLP approach compared favorable with traditional methods (Templeton & Guillemin, 1990).

Although the connectionist models have been proved useful in dealing with statistical data and although their performances in speech signal processing have shown to be somewhat superior to traditional methods, they have difficulty handling the sequential character of speech. In other words, they are not able to properly exploit time-varying input patterns in terms of a sequence of output classes. Therefore, some researchers are thinking of using a combination method in speech signal processing, which can incorporate the advantages of both connectionist models and traditional methods. Bourlard and Morgan reported a hybrid method which merges MLP and hidden Markov models for continuous speech recognition (Bourlard & Morgan, 1991).

In the present study, template matching and MLP are combined to identify speakers. In the template matching method, cepstral coefficients were selected as acoustic features, and an efficient DTW algorithm (Zhu & Cahill, 1991) was used to compare the feature vectors at equivalent points in time. A nearest neighbor rule was used to determine the identity of the speakers. In the combination approach, an unknown speaker's template was first compared with all the stored speakers' reference templates to choose a few candidates. An MLP method, in which formant parameters of vowels and diphthongs were chosen as features, was then used on these candidates to identify the the identity of the speaker. A two-layer feed-forward MLP is used in the MLP method. The perceptrons are trained by the back-propagation algorithm based on the criterion of minimizing root-mean-square error between the output of the MLP and the desired output (Hecht-Nielsen,1989).

SPEAKER IDENTIFICATION BY TEMPLATE MATCHING

Speech database

Two kinds of test utterances are used in the experiments. One is a phonetically-diverse isolated word utterance '1 8 m b', the other is an all-voiced continuous sentence 'I know when my lawyer is due'. The reasons of choosing these as test utterances are that they both contain vowels or diphthongs whose formants have been shown to be effective to speaker recognition, and they both contain not only vowels and consonants, but also nasals which contain speaker-specific information. The utterances were recorded by nine male speakers and four female speakers under normal noise in a laboratory. The recordings were made in two sessions separated by a number of weeks. Each session contained the thirteen speakers' four repetitions of the two utterances, i.e., $4 \times 2 \times 13 = 104$ utterances. Data were captured at an 8 KHz sampling rate with 12-bit precision, using μ-law companding to produce 8-bit data samples. The resulting data format was compatible with the U.S. standards for digital voice transmission over telephone lines.

Cepstrum extraction

Cepstral coefficients were extracted through linear prediction analysis:

1. The beginning and end points of the recorded speech utterances were detected. To the isolated word utterances, the silent intervals between two neighboring words were cut off. The endpoint detection was accomplished by means of an energy calculation. A conservative threshold was used in word boundary detection to ensure that all the speech intervals remained.
2. A high emphasis filter $(1-0.95Z^{-1})$ was applied to the delimited speech. A 32 ms (256-point) Hamming window was used on the emphasized speech every 8 ms.
3. The first to fourteenth LPC coefficients were extracted from each frame by the autocorrelation method. The cepstral coefficients were found from the LPC coefficients by a recursive formula (Sethuraman & Gowdy, 1989).

Cepstra were normalized over the duration of the entire utterance to reduce long-term intraspeaker spectral variability.

Speaker identification

The process of template matching was accomplished by using the efficient DTW algorithm which is particularly suitable for speaker recognition (Zhu & Cahill,1991). The reference template of each speaker's utterance was created by averaging the cepstra of the speaker's four repetitions in the first session. The feature vectors in the reference template were the average values of the four cepstral coefficients at equivalent points in time. The utterances in the second session were used as test templates. The Chebyshev distance was used as a measure of dissimilarity between the reference and test templates:

$$D(R(n),T(m)) = \sum_{i=1}^{K}(r_i(n) - t_i(m))^2, \tag{1}$$

where $D(R(n),T(m))$ is the local distance between the nth frame of the reference template $R(n)$ and the mth frame of the test template $T(m)$, $r_i(n)$ and $t_i(m)$ are the ith $(i = 1,2,...,K)$ elements of $R(n)$ and

$T(m)$ respectively, and K is the dimension of the feature vector, $K = 14$ in the present experiment.

The nearest neighbor rule was implemented for determining the identity of the unknown speaker. The unknown feature vector was compared with each of the reference vector for that test utterance, and the overall distance accumulated for the optimum warping path for each speaker was evaluated. The unknown speaker was identified as the reference speaker whose template had the minimum accumulated distance with the unknown feature vector.

SPEAKER IDENTIFICATION BY MULTILAYER PERCEPTRON

Formant parameters

1. By reading the spectrograms of the utterances and by listening the utterances, the vowel [i] and the diphthong [ei] from the isolated utterance and the diphthongs [ai] and [ɔi] from the continuous utterance were segmented.

2. The first to fourteenth LPC coefficients were extracted from every frame of the spotted utterances by the method described in the last section.

3. The first three formants $F1$,$F2$ and $F3$ were extracted by identifying the peaks on the LPC spectrum and by hand editing of the resulting candidates on the basis of estimated bandwidths, formant ranges and continuity.

4. By comparing the extracted formant tracks with typical formant values of vowels and diphthongs, the exact vowel or diphthong portions were spotted. Formant parameters were calculated within the portions. In our experiments, eighteen formant parameters were used to distinguish speakers. Table 1 lists the parameters and their descriptions. More parameters related with the second formants of the diphthongs were used, since the tracks of second formants in diphthongs are very efficient at identifying speakers. All these parameters have been shown to be efficient for speaker recognition (Goldstein, 1976).

Parameter	Description
AVEi1	average first formant of [i]
AVEi2	average second formant of [i]
AVEi3	average third formant of [i]
AVEai1	average first formant of [ai]
AVEai2	average second formant of [ai]
AVEai3	average third formant of [ai]
MAXai2	maximum second formant of [ai]
MSPai2	slope in middle portion of second formant of [ai]
AVEɔi1	average first formant of [ɔi]
AVEɔi2	average second formant of [ɔi]
AVEɔi3	average third formant of [ɔi]
MAXɔi2	maximum second formant of [ɔi]
MSPɔi2	slope in middle portion of second formant of [ɔi]
AVEei1	average first formant of [ei]
AVEei2	average second formant of [ei]
AVEei3	average third formant of [ei]
INLei2	initial frame of second formant of [ei]
MIDei2	middle frame of second formant of [ei]

Table 1: Formant parameters and their descriptions.

Multilayer perceptron

The MLP used in our experiments has eighteen inputs, eighteen units in a single hidden layer, and thirteen outputs. The inputs of the MLP were the scaled formant parameters of every speaker. The thirteen outputs corresponded to the thirteen speakers.

The network was trained by the back-propagation training algorithm based on the criterion of minimizing root-mean-square estimation error between the outputs of MLP and the desired outputs (Hecht-Nielsen,1989). The data of the first record session were used as training data. To each training pattern, the desired outputs had only one high signal (a value between 0.5 and 1), which corresponded to the speaker. The other desired outputs were set to a low value (between 0 and 0.5). The network was trained by all the speakers in turn. Training terminated when the average root-mean-square error of all the training data fell within an acceptable limit.

The data of the second record session were used to test the network. The formant parameters of the test utterances were input to the network, the highest output corresponded to the identified speaker.

EXPERIMENTS AND RESULTS

In our experiments, 104 utterances in the first record session were used as training data, and 104 utterances in the second session as test data. The test set is close, i.e., all test speakers are in a known set. Tables 2, 3 and 4 show confusion matrices of the template matching the isolated word utterance, the template matching the continuous sentence and the MLP approach respectively. In the tables, the speaker in column i is identified as the speaker in line j. The speaker 1, 2, ..., 9 are males, and the others are females.

	1	2	3	4	5	6	7	8	9	10	11	12	13
1	4	0	0	0	0	0	0	0	0	0	0	0	0
2	0	4	0	0	0	0	0	1	0	0	0	0	0
3	0	0	4	0	0	0	0	0	0	0	0	0	0
4	0	0	0	3	0	0	0	0	0	0	0	0	0
5	0	0	0	0	4	0	0	0	0	0	0	0	0
6	0	0	0	1	0	3	0	0	0	0	0	0	0
7	0	0	0	0	0	0	4	0	0	0	0	0	0
8	0	0	0	0	0	0	0	3	0	0	0	0	0
9	0	0	0	0	0	1	0	0	4	0	0	0	0
10	0	0	0	0	0	0	0	0	0	4	0	0	0
11	0	0	0	0	0	0	0	0	0	0	3	0	1
12	0	0	0	0	0	0	0	0	0	0	0	4	0
13	0	0	0	0	0	0	0	0	0	0	1	0	3

Table 2: Confusion matrix of the template matching the utterance "1 8 m b".

The overall identification rates in Tables 2, 3 and 4 are 90%, 92% and 85% respectively. That the identification rate of the MLP approach is a bit lower than that of the template matching approach does not necessarily mean that connectionist models are inferior to conventional models in speaker recognition. The reason for the results may be that only a few vowels and diphthongs were used to distinguish speakers in the MLP approach. Moreover, the formant parameters in the experiment have not been shown as the best parameters for speaker identification. (This will be done at a later work.) The female speakers are more difficult to identify since their higher pitches and the narrow band filtering affect the accuracy of the formant extraction.

It is worth noting that the two methods are partly complementary: the errors of the template matching approach are not all the same as those of the MLP approach. So these two methods can be combined to get a higher overall identification rate. To achieve this, the following decision process was implemented in the

experiment of the combination method. First of all, the minimum accumulated distances between the test template and all the reference templates were compared with a pre-determined threshold. If only one distance was less than the threshold, the test speaker then was identified as the speaker. If more than one distance fell within the threshold, all of the corresponding speakers were considered as candidates of the true speaker. If none of the distances was less than the threshold, the three speakers whose templates were closest to the test template were chosen as the candidates. Then, the MLP method was used on these candidates to identify the speaker. The final identification rate is 98%.

	1	2	3	4	5	6	7	8	9	10	11	12	13
1	4	0	0	0	0	0	0	0	0	0	0	0	0
2	0	4	0	0	0	0	0	0	0	0	0	0	0
3	0	0	4	0	0	0	0	0	0	0	0	0	0
4	0	0	0	3	0	0	0	0	0	0	0	0	0
5	0	0	0	0	4	0	0	0	0	0	0	0	0
6	0	0	0	1	0	3	0	0	0	0	0	0	0
7	0	0	0	0	0	0	4	0	0	0	0	0	0
8	0	0	0	0	0	0	0	4	0	0	0	0	0
9	0	0	0	0	0	1	0	0	4	0	0	0	0
10	0	0	0	0	0	0	0	0	0	4	0	0	0
11	0	0	0	0	0	0	0	0	0	0	3	0	1
12	0	0	0	0	0	0	0	0	0	0	0	4	0
13	0	0	0	0	0	0	0	0	0	0	1	0	3

Table 3: Confusion matrix of the template matching the utterance "I know when my lawyer is due".

	1	2	3	4	5	6	7	8	9	10	11	12	13
1	4	0	0	0	0	0	0	0	0	0	0	0	0
2	0	3	0	0	0	0	0	0	0	0	0	0	0
3	0	0	4	0	0	0	0	0	0	0	0	0	0
4	0	0	0	4	0	0	0	0	0	0	0	0	0
5	0	1	0	0	2	0	0	0	0	0	0	0	0
6	0	0	0	0	0	4	0	0	0	0	0	0	0
7	0	0	0	0	0	0	4	0	0	0	0	0	0
8	0	0	0	0	1	0	0	3	0	0	0	0	0
9	0	0	0	0	0	0	0	0	4	0	0	0	0
10	0	0	0	0	1	0	0	1	0	3	0	0	0
11	0	0	0	0	0	0	0	0	0	0	3	0	2
12	0	0	0	0	0	0	0	0	0	0	0	4	0
13	0	0	0	0	0	0	0	0	0	1	1	0	2

Table 4: Confusion matrix of the MLP approach.

CONCLUSION

The experiments described above have shown that the proposed combination approach is superior to either the template matching method or the MLP method in speaker identification. Although this work is restricted to text-dependent speaker identification for a small number of speakers, the results suggest that it will achieve better performances in speaker recognition by taking advantage of both connectionist models and convention models.

Bourlard, H. & Morgan, N. (1991) *Merging multilayer perceptrons and hidden Markov models: Some experiments in continuous speech recognition*, Neural Network Advances and Applications, (North-Holland: The Netherlands), 215-239.

Goldstein, U.G. (1976) *Speaker-identifying features based on formant tracks*, J. Acoust. Soc. Am. 59, 176-182.

Hecht-Nielsen, R. (1989) *Neurocomputing*, (Addison-Wesley).

Miles, M.J. & Guillemin, B.J. (1989) *Speaker recognition based on an analysis of vowel sounds*, IREE Int. Conf.: Digest of Papers, Melbourne, Australia, 120-123.

Naik, J.M. (1990) *Speaker verification: A tutorial*, IEEE Communications Magazine, Jan. 1990, 42-48.

Sethuraman, R. & Gowdy, J.N. (1989) *A cepstral based speaker recognition system*, Proc. Twenty First Southeast Symp. Syst. Theory, Tallahassee, FL, USA, 503-507.

Templeton, P.D. & Guillemin, B.J. (1990) *Speaker identification based on vowel sounds using neural networks*, Proc. Third Australian Int. Conf. on Speech Science and Technology, Melbourne, Australia, 280-285.

Zhu, X.Y. & Cahill, L.W. (1991) *An efficient dynamic time warping algorithm for matching two indefinite utterances*, Proc. IREE Int. Conf., Vol.1, Sydney, Australia, 293-296.