Field Evaluation of Text-Dependent Speaker Recognition in an Access Control Application

Harsh Gupta, Ville Hautamäki, Tomi Kinnunen and Pasi Fränti

Speech and Image Processing Unit,
Department of Computer Science,
University of Joensuu, Joensuu, Finland
{gupta, villeh, tkinnu, franti}@cs.joensuu.fi

Abstract

Vector quantization (VQ) is a widely used matching algorithm for text-independent speaker recognition. In this paper we study the use of text-dependent speaker recognition in practical access control application. We compared dynamic time warping (DTW) to VQ-based matching using text-dependent pass phrases. Our goal was to find out, how fixed phrase affects speaker recognition performance. We collected corpus of 21 speakers at the location of access control system and experimented with two different text-dependent scenarios: with speaker dependent phrases and with speaker independent phrases. In both cases, DTW outperforms VQ matching, or works similar. Also text-independent test were carried out.

1. Introduction

One of the advantages of speech as a biometric identifier is user convenience. Firstly, it is possible to give the identity claim using the same modality as the biometric sample itself [1]. Secondly, it has been proposed that in a fixed phrase system, the user could select himself/herself the pass phrase [2]. Having this in mind, one possible application of speaker recognition technology is access control into physical facilities where user convenience is to be preferred over security. For example in the hospital environment, doctors and nurses should have a priority access to elevators with minimal user interaction. Speaker recognition technology could be used in this case as a biometric access control method. Furthermore, security of the system is not nearly as important as user convenience of it.

Having these considerations in mind, traditional speaker verification scenario might not be the best choice for user-convenient applications. Instead, open-set speaker identification seems a more feasible choice (Fig. 2). In an open-set identification system, the unknown sample is compared with all the voice templates of the authorized persons. If the best-matching speaker's score exceeds a threshold, the user is accepted and otherwise rejected. This threshold is usually common for all users, but there is one reason for making it user-depended: some person(s) might be so called 'sheep' so that everyone (also non-authorized persons) is identified as being this person [3]. For this user, the threshold should be set higher in general.

In this study, we consider an access control system prototype implemented by our research group and installed in our lab door, see Fig. 1. In text-independent speaker recognition corpus simulations, low error rate is easy to achieve [4, 5, 6]. On the other hand, we have noticed in practice that low error rate is difficult to achieve in everyday environment with short test phrases. It is clear that when user convenience is more important than security, system should be able to verify the user with very a short phrase. Text-dependent speaker recognition achieves speaker recognition from short pass phrases. We study our access control system in text-dependent usage scenario. In text-dependent scenario, we compare the effect of using common phrase or speaker dependent phrases. Also we compare pattern-matching algorithm (vector quantization vs. dynamic time warping).

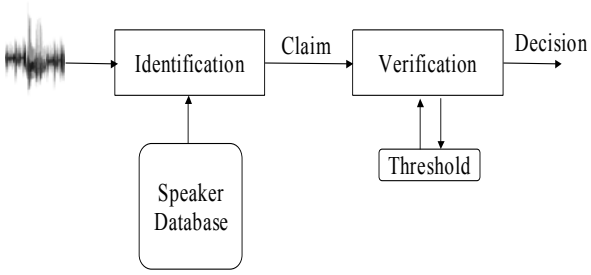

Figure 1: *Open-set speaker identification system* .

The rest of this paper is organized as follows: In Section 2 access control system is described and Section 3 describes in detail both DTW and VQ approaches. Section 4 presents the experimental setup and data collection. Section 5 gives experimental results and Section 6 gives conclusions.

2. Description of the system

The door to enter our lab has an electronic lock, connected to a PC via a Bus control unit and Door control unit. MFCC and VQ based speaker recognition software developed by our research group, which allows enrolling of speakers into the database and carrying out speaker recognition and verification, has been installed on the PC. There is a button connected to the PC near the door, on pressing which the software starts to record voice via the microphone near the door, and stops recording on pressing the button again. (Figure 2) This unknown speech sample is matched against enrolled speakers in the speaker database. Then identified speaker label is sent as a claim to the verification system, which decides, whether to open the door or not. If approved, the software sends a message to the Door Control Service running on the PC, which in turn opens the door lock via the Door Control Unit.

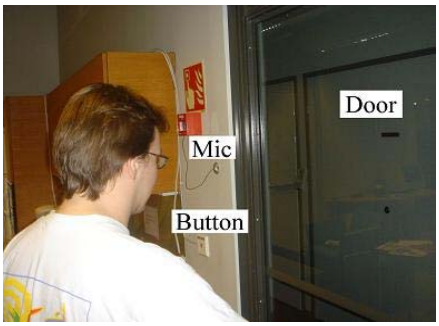

Figure 2: *Speaker recognition system in access control.*

3. Template Matching Methods

We applied both the dynamic time warping and vector quantization template matching methods. In both methods, template C_i is first created for each enrolled speaker *i* and then unknown sample *X* is matched against all templates. Matching outputs a dissimilarity score *D(.)*, from where speaker with minimum score is selected as the identified speaker.

$$Id = \arg\min_{i=i\ldots N} \{D(X, C_i)\}$$

Mel Frequency Cepstral Coefficients (MFCC) features were used to represent raw speech signal into compact but effective representation that is more stable and discriminative than the original signal [6]. First the pre-emphasized and windowed speech frame is converted into spectral domain by the Fast Fourier Transform (FFT). The magnitude spectrum is then smoothened by a bank of triangular bandpass filters that emulate the critical band processing of the human ear. Each of the bandpass filters computes a weighted average of that subband, which is then compressed by logarithm. The log compressed filer outputs are then decorrelated using the Discrete Cosine Transform (DCT). The zeroth cepstral coefficient is discarded since it depends on the intensity of the frame.

3.1. Dynamic Time Warping

Dynamic time warping uses the principle of *dynamic programming* [principle of optimality], in order to compute overall distortion between the two speech templates. Comparing the template with incoming speech might be achieved via a pairwise comparison of the feature vectors in each. The problem with this approach is that if constant window spacing is used, the length of the input and stored sequences is unlikely to be the same. Moreover, within a word, there will be variation in the length of individual phonemes. The matching process needs to compensate for length differences and take account of the non-linear nature of the length differences within the words.

The dynamic time warping algorithm achieves this goal, it finds an optimal alignment between two sequences of feature vectors, which allows for stretched and compressed sections of the sequence [7].

We can arrange the two sequences of observations on the sides of a grid with the unknown sequence on the bottom and the stored template up on the left. Both sequences start on the bottom left of the grid. Inside each cell we can place a distance measure comparing the corresponding elements of the two sequences.

Step 1 *Initialization*
D(1,1) = d(1,1)m B(1,1), for j = 2,....,M compute D(i,j) = ∞
Step 2 *Iteration*
for i = 2,..., N {
for j = 1,...,M compute {
$D(i, j) = \min_{1 \le p \le M} [D(i-1, p) + d(p, j)]$
$B(i, j) = \arg\min_{1 \le p \le M} [D(i-1, p) + d(p, j)]$ }}
Step 3 *Backtracking and Termination*
The optimal (minimum) distance is D(N,M) and the optimal path is $(s_1, s_2, ..., s_N)$ where $s_N = M$ and $s_i = B(i+1, s_{i+1}$, i = N-1, N-2, ..., 1

Figure 3: *The Dynamic Programming Algorithm [8]*

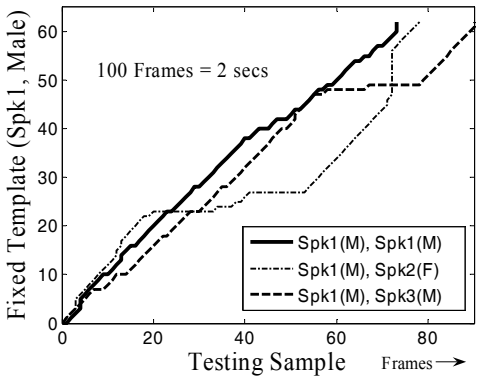

Figure 4: *Optimal DTW paths*

To find the best match between these two sequences we can find a path through the grid which minimizes the total distance between them. Figure 4 shows the optimal paths obtained by comparing the template of a speaker with his own testing sample and two other speakers, one male and other female. The DTW algorithm is designed to exploit some observations about the likely solution to make the comparison between sequences more efficient. Instead of finding all possible routes through the grid that satisfy these constraints, the DTW algorithm works by keeping track of the cost of the best path to each point in the grid. The purpose of DTW is to produce a warping function that minimizes the total distance between the respective points of these phrases. We used symmetric DTW with slope constraint = 0. The distances in DTW were calculated using Euclidean distance metric.

3.2. VQ Matching

In the VQ-based approach to speaker identification, the speaker models are formed by clustering the speaker's feature vectors into *K* disjoint clusters [4, 5, 9]. Each cluster is represented by a code vector c_i, which is the centroid (average vector) of the cluster. The resulting set of code vectors $\{c_1,.., c_k\}$ is called a codebook, and it serves as the model for the speaker. However, in these experiments we used raw feature set extracted from the training speech as a speaker model. This was done because speech segment was so short that clustering was not practical.

The matching function in the baseline VQ-based speaker recognition [9] is the quantization distortion between the two vector sets to be compared. Given a feature vector xi generated by the unknown speaker, and a codebook C = {c_1 …c_k }, the total quantization distortion D_Q is given by

$$D_Q(X,C) = \frac{1}{|X|}\sum_{x_i \in X} \min_{c_j \in C} d(x_i, C) \qquad (1)$$

where d(·,·) is a distance metric defined over the feature space. Typically Euclidean metric is used as the distance measure. The normalization factor 1/|X | is the same for all speakers, and therefore it does not change the order of the speakers in the matching result.

4. Experimental Setup

In order to analyze the effectiveness of both methods in the real world applications, all the samples were recorded by the Door Access Control System, in place in our lab at Speech and Image Processing Unit (SIPU), University of Joensuu. This allowed ambient noises and room acoustics to interfere with the recordings. Samples from 21 speakers were collected, out of which 7 were females and rest males, none of which were native English speakers. All the phrases were in English. One of the two training phrases was common to all speakers. This phrase was used in common phrase testing. The second phrase was unique for every speaker. In both cases, the same corresponding sentence was used for testing. The speaker-dependent phrase varied in length for every speaker, and no emphasis on the phonetic content of these phrases was laid. Common phrase was "*Every salt breeze comes from the sea*" and speaker dependent phrases were sentences from newspaper clippings (Appendix A). Recordings were done at the location of entrance door of our lab. All samples were recorded at 22 KHz, in 16-bit PCM WAV format. The recorded samples were manually aligned by removing the initial and trailing silence.

4.1. Feature Extraction

MFCC features have been used for evaluating the performance of both DTW and VQ-Matching. We used MFCC features of order 12, with window size = 30 ms. and window shift = 20 ms. Features were extracted using the command line automatic speaker identification software developed at our lab, hereafter referred to as "sprofiler". First the features were extracted for all the samples. Then VQ-matching based speaker identification results were evaluated using sprofiler and for DTW based results, code available at Helsinki University of Technology website was used.

5. Results

We evaluated the error rates on our corpus for both the text-dependent scenarios: speaker independent and speaker dependent. The results have been summarized in Table 1. The row corresponding to 'S1 – S2' quotes the results obtained when the voice samples obtained in first recording were used as testing samples and the samples collected in second recording were used as training samples. The next row corresponding to 'S2 – S1' in the first column quotes the results by interchanging the training and testing samples that were used in the previous test.

Table 1: *Identification error rates in speaker identification*

	Speaker Dependent		Speaker Indpendent	
Test -Train	DTW	VQ	DTW	VQ
S1 – S2	4.8	0	14.3	19.1
S2 – S1	0	19.1	19.1	28.6

As can be inferred from the table, DTW gives better or comparable results with VQ-matching for all the cases. Moreover, it can be seen that the samples collected when the user has become acquainted with the system and the fact that he/she is being recorded doesn't bother him/her anymore, serve as better training samples to compute the templates. Thus training samples should be as long as permissible by the application where speaker recognition system is being deployed, so that the length of the sample can compensate the variation in the speech pattern of the user.

5.1. Sample length

In order to assess the effect of the length of training and testing samples on speaker identification, we first aligned both the samples of the same speaker using DTW and then made 10 equal sections of the optimal alignment path found.. The feature vector files of the samples were copied from the beginning to the point indicated by the i^{th} section of the minimal path. Figure 5 summarizes the results obtained. The 'VQ(TI) Matching' plot gives the error rates of using VQ-Matching for Text-Independent scenario.

DTW gives much less error rates on smaller samples, in both speaker independent and speaker dependent scenarios, as per our expectations. This makes DTW fit for use in case of speaker identification using small pass phrases, where it outperforms VQ by a good amount. If samples are long enough then both VQ and DTW perform on a similar level, which is much better than the Text-Independent case.

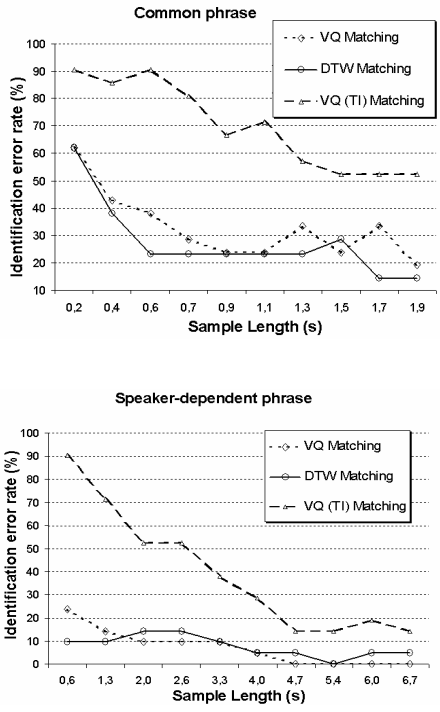

Figure 5: *Effect of sample length on VQ and DTW matching*

5.2. Verification

The verification results for are presented in Fig. 6 as a receiver operating curve (ROC) showing the genuine acceptance rate as a function of the false acceptance rate. The equal error rates are also summarized in Table 2.

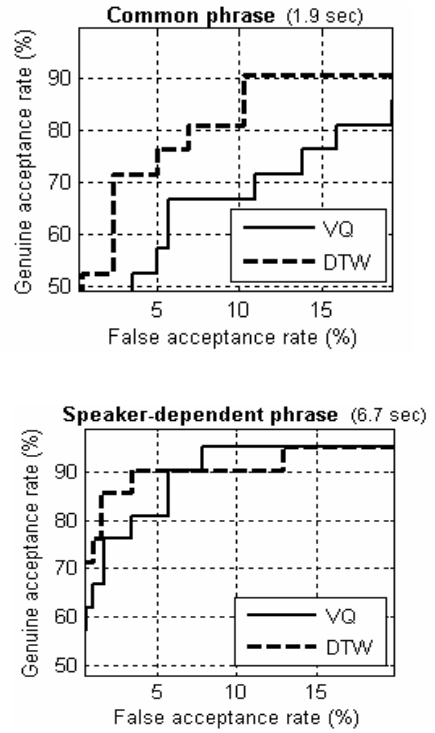

Figure 6: *ROC Curves for different scenarios*

Algorithm	Common Phrase	Speaker Dependent Phrase
DTW	10.2	9.5
VQ	19.0	7.9

Table 2: *Equal Error Rates*

For both the common and speaker-dependent phrase, DTW outperforms VQ. Especially for the common phrase in which the utterance is much shorter, the absolute difference in the both false acceptance and false rejection rate is about 5-10 %. For the speaker-dependent case with longer utterances, the differences are smaller, but DTW outperforms VQ at low false acceptance rates. On the other hand, in the user - convenient end of the tradeoff curve the differences are smaller. Based on these results, we suggest using DTW if the pass phrases are short or security (low false acceptance rate) are important. For longer pass phrases and user-convenience, the method does not play so important role.

6. Conclusions

In the view of results presented here, the DTW approach outperforms VQ if the pass phrases are short. For longer pass phrases the two approaches perform essentially similar. However, from the user convenience point of view, the pass phrase should be as short as possible, and thus we recommend to use DTW as the default method. In future, effects of the pass phrase and intersession variability should be studied. The results should be also validated on a larger database.

7. Acknowledgements

Work of Harsh Gupta was supported by the National Technology Agency of Finland (TEKES) (project "*Puhetekniikan uudet menetelmät ja sovellukset*", http://cs.joensuu.fi/pages/pums, TEKES dnro 8713103).

8. References

[1] Heck L. and Genoud D., "Combining Speaker and Speech Recognition Systems", Proc. Int. Conf. on Spoken Language Processing, p 1369-1372, 2002.

[2] Rosenberg, Aaron E. / Parthasarathy, S. (1997): "Speaker identification with user-selected password phrases", Proc. 5th European Conf. on Speech Communication and Technology, p 1371-1374, 1997

[3] Campbell, J.P. "Speaker Recognition: a Tutorial", Proceedings of the IEEE, Vol. 85, no. 9, 1997, p. 1437-1462

[4] Kinnunen, T., Karpov, E. and Fränti, P., "Real-Time Speaker Identification and Verification", IEEE Transactions on Speech and Audio Processing (accepted for publication).

[5] Saastamoinen, J., Karpov, E., Hautamäki, V. and Fränti, P.,"Accuracy of MFCC based speaker recognition in Series 60 device", EURASIP Journal on Applied Signal Processing, (accepted for publication).

[6] Kinnunen, T., Hautamäki, V., Fränti, P., ""Fusion of Spectral Feature Sets for Accurate Speaker Identification", Proc.9th International Conference Speech and Computer (SPECOM'2004), pp. 361-365, St. Petersburg, Russia, September 20-22, 2004

[7] Sakoe, H. and Chiba, S. (1978) 'Dynamic programming algorithm optimization for spoken word recognition,' IEEE Trans. ASSP, vol.26, no. 1,43-49.

[8] Huang X., Acero A., Hon H., (2001) 'Spoken language processing: a guide to theory, algorithm, and system development', Prentice-Hall, New Jersey, USA

[9] Soong, F.K., Rosenberg, A.E., Juang, B. H. and Rabiner, L.R., "A vector quantization approach to speaker recognition", AT & T Technical Journal, Vol. 66, p. 14-26, 1987.

Appendix A

- Finland's guards are threatening to start a new strike on the week of the Midsummer holiday.- According to a labour market advisor of the Central Organisation of Finnish Trade Unions, poor treatment of Estonians working in Finland is an everyday occurrence. - A report published in Sweden on Tuesday on the response of the country's officials to the aftermath of the Asian tsunami sharply criticises the response of Swedish officials to the disaster. - The decision by the Finnish Frontier Guard to stop inspecting passports of travellers is not a violation of the Schengen Treaty. - The change in conditions in the passenger traffic on the Baltic Sea is plain to read from the financial reports.- Though it is early days yet, the summer of 2005 seems frighteningly similar to the one before. - The Labour Court ruled on Monday that the tasks the Frontier Guard had been allocating to non-union personnel at crossing points could not be labelled safety work. - A fresh study indicates that Finnish elderly people tend to be more open-minded than senior citizens in other European countries. - Persistent rain on Saturday made a mess of the time-honoured tradition of new upper secondary graduates. Finnish pulp and paper mills will remain shut down for the time being. - An extensive Trafficking in Persons Report by the United States Department of State ranks Finland among countries which do not fully comply - The decision by the Finnish Frontier Guard to stop inspecting passports of travellers is a violation of the Schengen Treaty after all.- Fortum's share price has more than doubled in just over a year, and it had been expected that the state would take advantage of the situation and sell its extra stock. - The Finnish Parliament is to debate the proposed constitution for the European Union in the autum. - The EU has existed for decades without a constitution, and can certainly continue. - With the introduction of the euro, consumers' ability to perceive the cost of things has become weaker. - Another interesting point is that the nominal value of the currency seems to bear significance in relation to consumer price awareness - Employees of the Leaf sweets factory in Turku walked off their jobs again on Thursday afternoon. - In the future, an estimated 150,000 people will pass through Helsinki's new Kamppi Center every day. - In one of the stranger stories to come to light this week, two large stray dogs apparently scared passers-by. - Russia has denied that its military planes had violated Finnish airspace over the Gulf of Finland.