Speaker Identification Using Neural Networks

Hitoshi Ihara, Hiroyuki Kamata and Yoshihisa Ishida

Department of Electronics and Communication , Meiji University
1-1-1, Higashi-mita, Tama-ku, Kawasaki, 214 Japan

ABSTRACT - In this paper we describe a speaker identification system using neural networks. We have found that there are mainly individual characteristics in the mid frequency band on the spectrum. We use this frequency band on the spectrum and the fundamental frequency. We show that the individual characteristics are included in that frequency band and consider the practical speaker recognition system.

1. INTRODUCTION

The research of speaker identification which is a technique to recognize a speaker using only voice have been attained in many countries [1-10]. We hope that this technique will be able to open the door, pay for shopping, show some information and access to the computer remotely by using only voice, and is useful in information society in the future (for example : home automation system).

There is already a study on speaker identification using longtime averaged speech spectrum [1]. This study needs the speech data in which each sentence is about 10 seconds length. On the other hand, we have done some studies using dynamic pattern matching and vector quantization method with comparatively long length words and they have a good recognition rate [2][3]. However, these methods need long recognition time and a lot of memories to register the reference vectors when the words of subject increase.

We consider the speaker recognition system that use the general words with short length utterances, and has less memory and more shorter recognition time compared with dynamic pattern matching. So we take a method using neural networks. The neural networks require less resources or computational intricacy compared with the pattern matching method.

We have many studies that use low or high frequency band in the speaker recognition system [1][4]. We examine which part on the spectrum of speech wave mainly includes the individualities. As a result, we have found that there are mainly individualities in the mid frequency band on the spectrum by principal component analysis and standard deviation analysis using the vowels /a/, /i/, /u/, /e/ and /o/. We try to examine the speaker identification using this frequency band based on fundamental experiment.

2. INFORMATION OF SPEAKER INDIVIDUALITIES INVOLVED ON THE SPECTRUM ENVELOPE

The spectrum envelope is used in many studies on the speaker identification [1, 3-5]. However, as it is well known, the region on speech spectrum which includes the speaker individuality has not been proved yet.

Then we take 8 patterns for each vowel (/ a /, / i /, / u / , / e / , / o /) uttered by 8 speakers, and search for the differences of speech spectrum by the principal component analysis using these vowels. This analysis looks for the frequency band which is the most different part on the spectrum envelope in the same vowel. The principal axis presents the difference in spectrum envelope for individual vowel. Fig. 1 shows an example of the spectra for vowel / e /. Fig. 2 shows the eigenvalue of the principal axis for vowel / e /. Then, we calculate the standard deviation for the above vowel and show the result in Fig. 3.

As the result, it is found that the difference has concentrated in the mid frequency band of about 2~6kHz in common to 5 vowels. There are less eigenvalues of the principal axis in the low frequency.

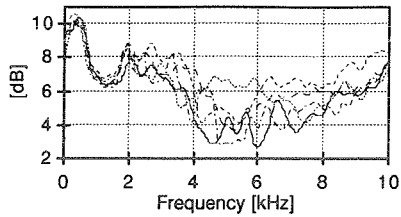

Fig. 1 The spectra of vowel / e /.

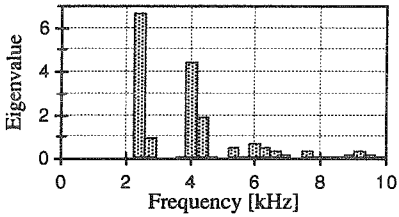

Fig. 2 The result of principal component analysis of vowel / e /.

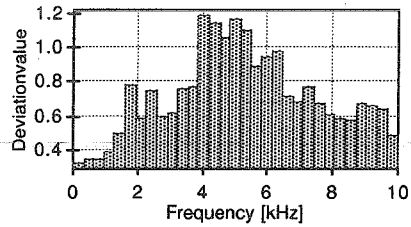

Fig. 3 The standard deviation of vowl / e /.

We consider that this frequency band includes the phonetic characteristics, because there are less differences. Furthermore, we can see that the difference of individual vocal tract shape appears in the mid frequency band on the spectrum envelope.

The speech production mechanism is represented by the product of the characteristics of glottis wave, vocal tract resonance and radiation. The characteristics of glottis wave with radiation tends to change easily by each speaker and the time when the speaker has uttered, compared with those of the vocal tract resonance [6].

So, we need to eliminate the glottis wave and radiation characteristics to obtain the spectrum exactly for speaker recognition. To eliminate these characteristics from the spectrum is able to compensate the change in the spectrun envelope by the speaker and utterance time. We ordinally use the pre-emphasis or adaptive inverse filter to eliminate these properties [6, 7].

We use the following filter in this study.

$$H(Z) = 1 - \varepsilon Z^{-1}$$

Usually, we use the coefficient ε with a fixed value. However, we decide this coefficient flexibly for each speaker and each utterance time.

3. EXPERIMENTS

3.1 Database

We record the spoken digits, /ZERO/, /ICHI/, /ROKU/ by 5 male speakers for speaker identification and they mean /ZERO/, /ONE/, /SIX/ in English. We record each digit 5 times to consider the change due to the difference of utterance time. We sample the speech data as follows. Utterance time 0 : 3 samples per day for each digit during continuous 3 days. Then ,Utterance time 1 : 4 samples, Utterance time 2 : 6 samples, Utterance time 3 : 4 samples and Utterance time 4 : 4 samples. The interval of each record time is a week. We record 27 utterances for each speaker. The speech data is sampled at 20 kHz.

3.2 Analysis

The sampled speech signal is automatically scanned its beginning and end points by fixed threshold level and we then get the spectrum and fundamental frequency. The threshold level separates the voice and unvoiced parts. For each frame, the time length of the sample data spans 25.6ms and frame interval is 5ms. In order to eliminate the influence of the glottis wave with radiation from the individual utterance, the adaptive pre-emphasis is used in the pre-process of the improved spectrum analysis. The spectrum and the fundamental frequency in the unvoiced part are neglected.

The speech signal is divided 8 segments to match the length of utterance. Then the spectra and the fundamental frequencies are averaged by each segment. We sample 14 points of spectrum data in 2~6kHz , and 2 points of spectrum data in 0~2kHz and 6~10kHz. Each segment is represented by the sampled values of 17 points including the fundamental frequency. Thus, 17x8 parameters are obtained from each speech signal. Then, the parameters are normalized by the dispersion value of themselves.

This operation is effective to compensate the variation of parameters for each speaker so as to bring each parameter close to its normal distribution [3, 8].

We make use of three-layered networks with 136 units in the input layer, 50 units in the hidden layer and 5 units in the output layer corresponding to 5 male speakers.

The networks are trained by the momentum method based on the well-known back-propagation algorithm.

3.3 EXPERIMENTAL CONDITIONS

3.3.1. Experiment 1 : The variation of recognition rate due the difference of utterance time

The training patterns to neural networks are 3 patterns in Utterance time 0 for each speaker. We examine the speaker identification using the networks trained by those patterns and observe the variation of recognition rate due to the difference of utterance time. We use the frequency band of 2~6kHz in this experiment. Furthermore, we try to observe the variation of recognition rate in the case of using 2 or 3 combined words.

3.3.2. The experiment 2 : In the case of changing the frequency band

There is a possibility that the speaker individualities appear in both of 0~2kHz and 6~10kHz. So we prepare some frequency bands as follows in addition to the frequency band of 2~6kHz; Frequency band 1 : 0~6kHz, Frequency band 2 : 1~6kHz, Frequency band 3 : 2~6kHz (base band), Frequency band 4 : 2~7kHz and Frequency band 5 : 2~8kHz.

4. THE EXPERIMENTAL RESULT

4.1. The result of experiment 1

Fig. 4 shows the variation of the average value in recognition rates for 5 speakers with respect to utterance time. The recognition rate changes for each word and for utterance time. The speaker recognition rate is 84.4% in case of using /ZERO/, 87.9% using /ICHI/, 84.2% using /SIX/ in the average for all speakers and all utterance times. We can see that the average value of the speaker identification rate is about 85% in this system.

Fig. 5 shows the variation of the speaker identification rate for each speaker due to utterance time. We can not get the precise result of presenting the relation between the speaker identification rate and utterance time because of using only sampled data for 4 weeks. We consider that if we observe the periodical variation of speech, we can obtain the speaker identification rate independent of utterance time.

On the other hand, Fig. 6 shows the speaker identification rate using some combined words. We see that the speaker identification rate is improved when we have used some combined words. The recognition rate is 84.8% in the average using only a word, in the case of using two words 93.6% in the average, and in the case of using 3 words 94.6%.

4.2 The result of experiment 2

We get the best recognition rate in the case of using the frequency band 3 : 2~6kHz form Fig. 7. The recognition rate tends to be not so much increasing as decreasing even if we consider the frequency band 0~2kHz and 6~10kHz. We consider that the information of speaker individualities comes to be rather vagueness because of including the frequency band of less individualities.

5. CONCLUSION

From the experiment 1, we get the result of the speaker identification at about 85% in the case of using a word. Then, in the case of using 3 words, the recognition rate increases about 95%. We consider that the proposed method is effective to the speaker recognition systems such as the security system of the door and the bank card which uses the password number.

From the result of the experiment 2, it shows that the information of speaker individualities is

included mainly in the frequency band of 2~6kHz.

However, this research examines only 5 speakers and 3 words. It is not obvious whether we can get the same result in the case of using more speakers and words.

REFERENCES

[1] S. Furui, F. Itakura and S. Sato, (1972), *Talker recognition by the longtime averaged speech spectrum*, IEICE, Vol. 55-A, No.10, pp.549-556.

[2] S. Furui, (1981), *Comparison of Speaker recognition methods using statical features and dynamic features*, IEEE, Trans. ASSP, vol. ASSP-29, NO.3, pp.342-350.

[3] T. Matui, S. Furi, (1992), *Text-independent speaker recognition using vocal tract and pitch information*, IEICE, Vol. J 75-A, No.4, pp.703-709.

[4] S. Hayakawa and F. Itakura, (1993), *Speaker recognition using information in the higher frequency band*, TECHNICAL REPORT OF IEICE, SP93-71, pp.55-61.

[5] B. S. Atal, (1974), *Effectiveness of linear prediction characteristics of speech wave for automatic speaker identification and verification*, J. Acoust. Soc. Am., Vol. 55, No. 6, pp.1304-1312.

[6] S. Furui, (1974), *An analysis of long-term variation of feature parameters of speech and its application to talker recognition*, IEICE, Vol. 57-A, No. 12, pp.880-887.

[7] H. Kamata, Y. Ishida and Y. Ogawa, (1990), *High speed estimation of vocal area function using digital signal processor*, T. IEE Japan, Vol. 110-D, No. 7, pp.773-780.

[8] H. Noda and T. Osanai, (1990), *On the relation between the number of speakers and the reliability of recognition rate in speaker recognition*, IEICE, Vol. J 73-A, No. 4, pp.717-724.

[9] S. Furui, (1982), *Speaker recognition by statical features of cepstral parameters*, IEICE, Vol. J 65-A, No. 2, pp.183-190.

[10] S. Furui, (1986), *Research on individuality features in speech wave and automatic speaker recognition techniques*, Speech Communication, Vol. 5, No. 2, pp.183-197.

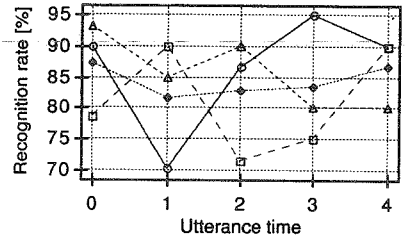

Fig. 4 The recognition rate for each words.
(O[0],Δ[1],□[6],◆the ave. of 3 words)

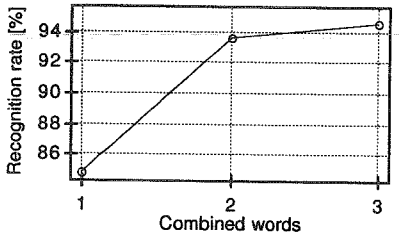

Fig. 6 The recognition rate in the case of using some combined words.

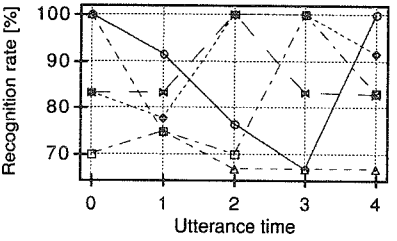

Fig. 5 The recognition rate for each speakers.
(mark [speaker] ; O[a], ◇[i], Δ[h], □[k], ⋈[m])

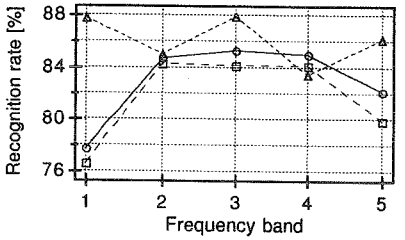

Fig. 7 The recognition rate in the case of using each frequency band.
(O[0],Δ[1],□[6])

THE EFFICACY OF COHORT NORMALISATION IN A SPEAKER VERIFICATION TASK UNDER DIFFERENT TYPES OF SPEECH SIGNAL VARIANCE

J.Bruce Millar, Fangxin Chen and Michael Wagner.

TRUST Project
Research School of Information Sciences and Engineering
Australian National University

ABSTRACT

This paper examines the influence of three types of speech signal variance on the performance of a text-independent speaker verification system and the efficacy of cohort normalisation as a technique to compensate for this influence. The three forms of variance comprise that generated by repetition of utterances over time, the addition of extraneous noise to test utterances, and the inclusion of test utterances which use phonemic sequences that do not occur in the training data. A statistical analysis of the results for a gender-balanced set of 20 client/impostor speakers and an independent population of 25 cohort speakers is presented. The results indicate that conventional cohort normalisation is moderately successful in combatting repetition variance and phonetic variance but gives no significant improvement in the presence of moderate levels of noise. A hybrid form of cohort normalisation is shown to combat the former two variance types very effectively and to provide a weakly significant improvement for moderate levels of noise induced variance.

INTRODUCTION

One of the major problems of speaker verification is the setting of thresholds against which to test measures of similarity or difference. The use of a fixed threshold is not robust under normal variance experienced in the capture of incoming speech in practical systems. Cohort normalisation has been shown to produce improved performance for speaker verification when applied to speech which has been captured under different conditions to those existing when the speaker models were created.

The basic idea is attributed to Higgins et al.(1991), and has been applied in other research on speaker verification (Rosenberg et al., 1992; Matsui and Furui, 1994). The TRUST project is exploring the use of this technique and has already highlighted one of its limitations (Chen et al., 1994a). A further paper in these proceedings (Chen et al., 1994b) extends our exploration of this technique.

The present paper aims to differentiate between the benefits of cohort normalisation when confronted with different types of speech signal variance. We examine the impact of three distinct forms of variance. The first form of variance is that produced by simple repetition of utterances by a speaker where the repetitions are spread across many days. The second form of variance is that produced when test speech has an ambient acoustic background which is different to that under which the speaker models were trained. Additive noise representing the presence of many speakers talking in the background is used in order to simulate this situation. The third form of variance is that produced by the use of phonetic sequences which were not a part of the training data from which the speaker models were created. Data fitting this specification were included in the initial TRUST data corpus (Millar et al., 1994).

While the first form of variance has been explicitly tested by others, the second and third forms of variance, although acknowledged have not been specifically explored. The aim of this work is to generate an effectiveness profile for cohort normalisation when challenged by explicitly controlled forms of speech variance.

The three forms of variance included in this paper are indicative of major problems facing speaker verification. Change in the client over time requires frequent updating of models unless a robust normalisation

technique can be used. The presence of extraneous noise which is different to that during the most recent enrolment is a constant hazard which cannot always be controlled for. The use of phonetic sequences which did not occur in training will occur due to speaking errors, changes in speaker physiology, and maybe minor task variations.

THE BASIC METHOD

The basic concept behind cohort normalisation is to provide for each client speaker a 'cohort' of similar speakers whose speaker models provide a local environment within which distance measurements to incoming test speech can be normalised. The cohort of similar speakers is selected by measuring the distance of prospective cohort members from the speaker model of the client speaker. The group of N speakers from a 'cohort set' who are closest to the client speaker are chosen to be members of that speaker cohort. This criterion may not be optimal (Chen et al., 1994b) but was the standard criterion in use at the time this work was performed.

The normalisation process involves the measurement of the distances between an incoming speech sample and the speaker models of the client and the members of the client's cohort. The unnormalised 'client distance' is normalised by subtracting some statistic of the set of cohort distances from the client distance. Our simple cohort normalised score is given by:-

$$D(norm) = D(client) - Minimum\ [D(cohort)]$$

This technique makes the assumption that the introduction of variance in the testing signal will cause a shift in the analytic space in which the speakers are modeled for the test utterance concerned. Thus the boundary between utterances which come from the client and those which come from other speakers will shift also. Verification that is based on a simple 'distance threshold' will be disturbed by such a shift. The 'cohort normalisation' formulation assumes that the variance-induced shift will vary the distance to the client model and to the models of similar speakers by a similar amount. A hybrid cohort normalised score is obtained by setting 'D(norm)' to a very large value when 'D(client)' exceeds a preset but large threshold lying well outside the clients normal speech range (Chen et al., 1994a).

THE DATA CORPUS FOR THESE EXPERIMENTS

In these experiments the initial TRUST data corpus (Millar et al., 1994) was used in the following way. Firstly 10 male and 10 female speakers were randomly selected to act as clients. A cohort of 5 similar speakers was chosen for each speaker from among an additional set of 25 speakers. The remaining 19 client speakers in the data corpus were used as impostors.

REPETITION VARIANCE

Repetition variance occurs when the same utterance is produced on successive occasions. Zhu et al. (1994) have shown that for the data from the initial TRUST corpus (Millar et al., 1994) the speaker identification performance of a vector quantisation (VQ) model deteriorates in a way that is dependent on the time interval between the production of the material on which the model was based and the production against which the model was tested. This is entirely consistent with earlier reports on the temporal variability of speaker characteristics (e.g. Furui, 1974; Soong et al., 1987)

In the current experiment VQ models were produced from the 5 repetitions of the corpus recorded in the first recording session (session A) for the 20 client speakers. Testing data comprised client data from the second (B) and third (C) sessions which were recorded at approximately one week intervals following the first session. Each session comprised 5 repetitions of the corpus. Impostor data was drawn from the utterances of the 19 other speakers in the client-set during sessions B and C.

The performance of each client model is expressed as an equal error rate (EER) percentage for each test session. The experiment was then repeated using simple and hybrid cohort normalisation providing results for both methods in the form of EER percentage for each client. These results are reported in table 1.