A STUDY OF INTRA- AND INTER-SPEAKER VARIABILITY IN VOICES OF TWINS FOR SPEAKER VERIFICATION

M. Mehdi Homayounpour(, **), Gérard Chollet(+)*
(*)CNRS/URA 1027, 19 rue des Bernardins, 75005, Paris, France
(**) Amirkabir University of technology, Hafez Street, Tehran, Iran

(+) IDIAP, C.P. 592, 1920, Martigny, Switzerland

ABSTRACT

This paper deals with the problem caused with similar voices like the voices of identical twins for text independent speaker verification. Three approaches to speaker verification were experimented: i) by human listeners, ii) by comparison of long-term spectra, and iii) by automatic methods [1]. A twin identification test was also conducted. Speaker verification experiments were achieved using LVQ3 and a Second Order Statistical Measure (SOSM). The results show that our automatic speaker verification systems discriminate the voices of identical twins worse than listeners familiar with them. It may be explained by the fact that twins relatives and their friends have received much more speech material for training than our automatic systems.

1. INTRODUCTION

Speaker verification algorithms perform well under controlled conditions, but their performance usually decreases when a user is recorded in other conditions or when he/she is in an emotional or pathological state, or when an impostor, an imitator or a person with a similar voice tries to be verified in his/her place. Twin brothers or sisters have similar voices in most cases. Rosenberg [4] and Cohen *et al.* [6] reported on speaker verification and identification experiments on voices of twins. Rosenberg did experiments with a single pair of twins. In his experiments, his automatic system performed better than human listeners. Cohen found that Cepstra and delta Cepstra yield adequate separation of voices of twins in a speaker identification task. Our experiments concern text independent speaker verification with 11 pairs of identical twins and siblings. These complementary experiments are described in sections 3 and 4. Section 2 specifies the content and recording conditions of the data base used for these experiments. Section 5 compares the results of speaker verification experiments done by human listeners and automatic systems.

2. THE TWIN DATA BASE

A telephone data base was recorded including recordings of 45 speakers consisting of 9 pairs of identical twins (8 males and 10 females) with similar voices, and 27 other speakers (13 males and 14 females) including 4 non-twin siblings. Each twin or sibling spoke for a total of 24 to 30 minutes in three sessions conducted with at least one week interval between sessions. In each session subjects were asked to read three different texts of one page. The speakers called from their office or from their home. Subjects were recorded over the telephone using an OROS AU32 PC-board at 16 bits linear form, 8 kHz sampling frequency.

3. LISTENING TESTS

For the aural method [5], listeners heard pairs of stimuli (55 pairs of 6s stimulus) extracted from the twin data base and decided whether they belonged to the same speaker or not. Two tests were conducted: In test I, there was no pair where both stimuli belonged to a twin pair, while test II included only pairs of stimuli belonging to twins or siblings. Test I was common for all the listeners but test II was different for the pairs of twin. Listeners were familiar or not with the twins or siblings.

Listening tests were conducted for the following purposes:

i) Is it an easy task for the human listeners to discriminate twins? What is the decrease of performance on twins.

ii) Is there a large difference in speaker verification performance when the listeners are familiar or not with the twins?

iii) Are the results of speaker verification by human listeners comparable to those of automatic systems on a twin data base?

In a further test (test III), family members of the twins were asked to listen at each time to a 6s stimulus of one of the twins and to identify him/her by using their a priori knowledge of the twins' voices. The result of this test and test II can serve to verify the hypothesis that when a listener is familiar with the twins (test III), he/she provide a smaller verification error rate (test II). Table 1, present the results of Test I.

Table 1- Results of speaker verification listening tests for test I and test II with Listeners Familiar With Twins (LFWT) and Listeners Not Familiar With Twins (LNFWT). FA and FR are False Acceptance and False Rejection error rates respectively. MER=(FA+FR)/2.

		FR(%)	FA(%)	MER(%)
LFWT	Test I	17.0	14.3	15.6
	Test II	20.2	16.4	18.3
LNFWT	Test I	16.8	14.6	15.7
	Test II	29.4	22.8	26.1

A 8.2% twin identification error rate (test III) was obtained for listeners familiar with the twins. It is much lower than the MER (18.3%) of test II for LFWT. Table 1 shows no bias in our population of listeners since identical results for LFWT and LNFWT are obtained on test I. On the contrary a highly significant difference is found between LFWT and LNFWT on test II. The error rates increase slightly from test I to test II when the listeners are familiar with the twins, while this increase is much more significant for LNFWT. A detailed observation of listening test results showed that when listeners are familiar with the twins, the twins identification error rate is directly proportional to twins speaker verification error rate. These results will be compared to those of our automatic approach in section 5.

4. AUTOMATIC APPROACH

Long-Term Spectra (LTS) and two automatic systems were developed and used for speaker verification. The automatic systems are based on a LVQ3 supervised neural net algorithm [4] and a SOSM measure[2]. 22 subjects comprising 18 twins and 4 siblings were considered as clients and 23 other subjects are impostors for our experiments with LVQ3 and SOSM.

4.1. Speech Analysis

Long silences were first removed. The recordings were pre-emphasised with a first order filter with transfer function of $1-0.95z^{-1}$. Each analysis frame of 30ms was multiplied by a Hamming window that was shifted by 15 ms. A vector of length 24 was retained which comprised 12 LPCC and 12 ΔLPCC [3]. Cepstral coefficients were normalised by subtracting from the cepstral coefficients their averages over the duration of the entire telephone call. This removes any fixed frequency-response distortion introduced by the transmission system.

The ΔLPCC coefficients represent the slope of the time-function of each coefficient in the cepstral vector; so it reflects the transitional information in speech signal. The regression slope is computed over 135 ms. Each coefficient in the feature vector was weighed by the reciprocal of its standard deviation obtained using 2s of training speech from each of the 22 clients.

4.2. Long-term Spectra

The identical twins have an identical, or at least very similar anatomy. So the speech differences between them is more related to their speech habits. This explains why most of our twins showed very close LTS when they were recorded over the same telephone line. LTS was very different when twins were recorded over different telephone line or handsets. Therefor LTS was rejected as a relevant feature to distinguish between twins.

4.3. LVQ3

Two speaker verification tests were conducted using a LVQ3 method adapted for speaker verification [1]. This technique allows to take other speakers into account during the training phase. A codebook for client i, contains three classes: one specifies client i, one for non-client i, and a class of noise and silence. The training data (reference vectors) for each client is obtained using 13.5s of speech from client i, 13.5s of speech from other clients having the same sex and 3.8s

of data representing silence, background and respiration noise. For a client i, the initial codebook contains 160 codes: 64 codes representing the class of client i, 64 codes representing the class of non-client i, and 32 codes representing noise. The initial codes were obtained by the classical LBG vector quantization algorithm using training data and were then tuned with the LVQ3 algorithm as explained in [1]. In the verification phase, the feature vector of a test utterance was compared to all vectors in the codebook and the code label of codebook-vector with the smallest distance to this feature vector was selected. This procedure was repeated for all feature vectors in the test utterance. A verification score was obtained which is equal to the number of testing vectors classified with the label I divided by the total number of vectors in test utterance minus the vectors classified as silence or noise. A speaker was accepted if his/her verification score was higher than a decision threshold, otherwise he/she was rejected.

Two experiments were conducted. They differ in the training phase. In the first experiment (1), training of a model for client i was done with data from any client of the same sex other than i. In the second experiment (2) training was done with 4 closest clients to i excluding twin or sibling. Verification tests were conducted with identical protocols for the two experiments:

x-a: tests on impostors

x-b: tests on twins and siblings.

where x reflects differences in training (x=1, 2). A test utterance duration of 6 seconds was used to conform with human listeners test conditions. The tests on impostors (test x-a) corresponds to protocol I of the listening tests while the tests on twins (tests x-b) is closer to protocol II of the listening tests. The FR obtained from the listening tests is applied to the FR/FA Receiver Operating characteristics Curve (ROC) of each client to find the corresponding FA. The mean of FR and this FA is considered as the error rate for this client (MER1). Similarly the FA of listening test is used to find the corresponding FR error rate and their average (MER2) is averaged with MER1 to find the Mean Error Rate (MER) for this client. The average of total error rates of all twins and siblings for the two sets of experiments is given in table 2. The speaker verification error rates are also presented by Equal Error Rate (EER).

Table 2. Results of speaker verification tests by LVQ3 method (experiments 1 and 2).

	MER	EER
1-a	18.0	13.1
1-b	30.0	30.3
2-a	19.6	21.9
2-b	31.1	34.6

4.4. SOSM

Another set of experiments were done with a SOSM technique [2]. The training speech data of the client i was used to compute a covariance matrix X for this speaker. A weighted symmetric sphericity measure $\mu(X, Y)$ is defined between a test covariance matrix Y and the reference covariance matrix X as the quantity:

$\mu_{SPH\ sym}(X, Y) = A+B$

where:

$A=\rho_{mn} \cdot \log(tr(YX-1)) + \rho_{nm} \cdot \log(tr(XY-1))$

$B= - \backslash F(1;m) \cdot (\rho_{mn} - \rho_{nm}) \cdot \log [\backslash F(det(Y); det(X))] - \log(m)$

with: $\rho_{mn} = \backslash F(m; m+n)$ and $\rho_{nm} = \backslash F(n; m+n)$

where m represents the number of training vectors and n the number of test vectors. For each client an individual covariance matrix was obtained using the same size of training speech material as used for training the LVQ3 models. Table 3 provides the results of the MER error rates obtained for experiments 3-a and 3-b

.

Table 3-Results of speaker verification test by SOSM method (experiment 3).

	MER	EER
3-a	13.5	8.7
3-b	30.0	28.5

4.4. LVQ3/SOSM

SOSM performs slightly better than LVQ3 on the protocol a. No significant difference is found on the protocol b. Both LVQ3 and SOSM show an increase in the verification error rate when a client's twin is considered as an impostor (tests x-b) compared to the case where speakers (non-clients) are impostors (tests x-a). The performance of our automatic systems degrades when a twin brother or sister tries to be verified in his/her place.

This decrease in performance is more important for SOSM method. A comparison of the results of experiment 1 and 2 for LVQ3 shows that when a larger number of speakers are taken into account for training a codebook for a client, a better speaker verification result can be obtained.

5. MACHINE vs. HUMAN

A comparison of Tables 1, 2 and 3, shows that neither human listeners nor our automatic systems are robust against voices of identical twins. Our automatic systems and listeners not familiar with the twins have about the same ability to discriminate between identical twins. The performance of human listeners didn't decrease significantly from test I to test II when the listeners are familiar with the twins. Our automatic systems behave in a way similar to listeners not familiar with the twins. The MER error rates which are obtained by taking into account the listening tests are higher than EER for both systems SOSM and LVQ3 methods. It shows that human listeners familiar with the twins and the two automatic systems studied here present different ROC (Receiver Operating Curve) characteristics.

6. CONCLUSION

Listening tests on twin voices showed generally an augmentation of false acceptance error rate for listeners not-knowing the twins and a smaller increase for listeners being member of the family or friends of twins. Human listeners familiar with twins may proceed with a first level of identification prior to discrimination. Long-term spectrum of speech was not found to be a relevant feature to discriminate between twins. Automatic speaker verification systems use only low level features which are related to the acoustic aspects of speech. The spectral representations of speech such as Cepstrum and delta Cepstrum parameters can not capture the behavioural differences between the twins. So a speaker verification system may take into consideration those features which represent the behavioural characteristics of a speaker to be more robust against the twins with similar voices. More efficient features and/or training procedures remain to be discovered to match the performance of listeners familiar with twins. But, of course, it should be noticed that twin relatives and friends have received much more speech material for training than our automatic systems.

ACKNOWLEDGEMENT

The authors would like to thank the twins and the other speakers and listeners who participated in data base recordings and listening tests.

REFERENCES

[1] M. M. Homayounpour, G. Chollet (1995) , "Neural Net Approaches to speaker Verification: Comparison with Second Order Statistic Measures",, *ICASSP'95.*

[2] F. Bimbot, L. Mathan (1994), "Second Order Statistical Measures for Text-Independent Speaker Identification", *ESCA Workshop on Speaker Recognition, Identification, and Verification*, pp. 51-54.

[3] M. M. Homayounpour, G. Chollet (1994), "A comparison of some relevant parametric representations for speaker verification", *ESCA Workshop on Automatic Speaker Recognition, Identification and Verification*, pp. 185-188.

[4] A. E. Rosenberg (1976), "Automatic Speaker Verification: A Review", *Proc. IEEE*, Vol. 64, pp. 475-487.

[5] Homayounpour, M. M., J. Ph. Goldman, G. Chollet, and J. Vaissière (, 1993), "Performance Comparison of Machine and Human Speaker Verification", *EUROSPEECH 93*, p. 2295-2298.

[6] A. Cohen, T. Vaich (1994), "On the Identification of Twins by Their Voices", *ESCA Workshop on speaker recognition, identification, and verification*, pp. 213-216.

[7] T. Kohonen (1990), "The Self Organizing Map", *Proceedings IEEE*, Vol. 78, pp. 1464-1480.