

SPEAKER VERIFICATION FROM A PROJECTION MATRIX

5

BACKGROUND OF THE INVENTION

10 The invention relates to a method for generating a verification projection matrix, a method for speaker verification, a device for generating a verification projection matrix, a speaker verification system, a computer readable storage medium, and a computer program element

15

In many applications, it is necessary to verify the identity of a speaker, i.e. an unknown person.

There have been proposed several verification methods using
20 biometric material of a person to be identified, e.g. the handwriting of a person, a person's eye structure, the fingerprint or even the speech.

When using speech signals for verification, the unknown
25 speaker who claims an identity speaks a predefined utterance, in the following denoted as speech signal, into the speaker verification system via a microphone. After having detected the speech signal, the speaker verification system usually performs a digitization of the speech signal. The speech
30 signal is subsequently divided into overlapping frames. The frames are time-aligned with a reference speech signal, spoken into the system during a training mode of the system by the true speaker, i.e. by the speaker whose identity is claimed by the unknown speaker. For each reference-aligned
35 frame, usually a feature vector is generated with specific predefined features which are chosen to be used in the verification process.

There are several approaches to use a so called principal component analysis (PCA) in the field of pattern recognition in general, speaker verification being a specific topic thereof.

US 5,054,083 discloses a speaker verification system, wherein the system is trained using utterances from true speakers and from impostors. The utterances are stored in a speaker data base. After having time-aligned the utterances, a diagonalized "in-class" covariance matrix is calculated by first computing the covariance matrices of the utterances from each true speaker. The covariance matrices are pooled together and the resulting "in-class" covariance matrix is diagonalized, thereby generating the diagonalized "in-class" covariance matrix. Furthermore, an "inter-class" covariance matrix is computed by pooling the covariance matrices for all successful impostors (i.e. for all utterances of an impostor, where the result of the verification was, that it was an utterance from a true speaker) of each true speaker over all true speakers. The final transformation matrix is computed by multiplying the diagonalized "in-class" covariance matrix with the "inter-class" covariance matrix and diagonalizing the resulting matrix. Using this matrix, the speaker verification for a new speech signal input by a speaker to be verified is performed.

However, the known speaker verification system often does not perform up to expectation under the noisy conditions.

SUMMARY OF THE INVENTION

Thus, it is a first object of the present invention to provide a method for generating a verification projection matrix used for verification of an unknown speaker which leads to improved speaker verification results especially

under noisy conditions, when the verification projection matrix is used in speaker verification.

To achieve the first object, according to a first aspect of the invention, a method for generating a verification projection matrix used for verification of an unknown speaker is provided. According to the invention, a feature vector for each frame of each speech signal out of a plurality of time-aligned speech signals j ($j = 1, \dots, m$) by the true speaker, is provided. Furthermore, an impostor feature vector for each frame of each speech signal out of a plurality of digitized time-aligned speech signals k ($k = 1, \dots, s$) by several impostors, is provided.

In other words, the true speaker speaks the predefined utterance into the speaker verification system m times during a training mode, thereby generating m speech signals. The speech signals are preprocessed, i.e. subjected to e.g. digitization, filtering, dividing into frames, and then one feature vector is generated for each frame comprising a predefined number of n features, i.e. the feature vector has dimension n . The speech signals are then time-aligned to a reference speech signal of the true speaker. After time-alignment, the m speech signals all comprise r reference-aligned frames, where r is the total number of frames of the reference speech signal. The time-alignment of the frames is necessary for making a comparison between the feature vectors possible. Furthermore, an arbitrarily number of impostors speak the same utterance into the speaker verification system f times during the training mode, thereby generating s speech signals which are pre-processed in the same way and which are time-aligned to the reference speech signal.

For all feature vectors of the i^{th} reference-aligned frame, there are executed the following steps in order to generate the verification projection matrix for the respective i^{th} reference-aligned frame:

In a first step, a principal component analysis on the feature vectors of the i^{th} reference-aligned frame of the true speaker is performed, thereby generating a first set of
5 eigenvectors, which form a first principal component matrix, and a first set of eigenvalues $\lambda_{i,j}$. Subsequently, a first matrix is selected from the first principal component matrix, wherein the first matrix comprises at least a part of the first set of eigenvectors associated with small eigenvalues
10 $\lambda_{i,j}$. Thus, using the first matrix, a first projection matrix may be generated by multiplying the first matrix with the transpose of the first matrix. Note that in the space defined by the first projection matrix, the feature vector of the true speaker has a small intra-speaker distance.

15 In a further step, the impostor feature vectors of the i^{th} reference-aligned frame are projected into the first space using the first projection matrix, thereby generating projected impostor feature vectors. Subsequently, a principal
20 component analysis on the projected impostor feature vectors of the i^{th} reference-aligned frame is performed, thereby generating a second set of eigenvectors, which form a second principal component matrix, and a second set of eigenvalues $\lambda'_{i,j}$.

25 In a last step, a second matrix is obtained as the verification projection matrix from the second principal component matrix, wherein the second matrix comprises at least a part of the second set of eigenvectors associated
30 with large eigenvalues $\lambda'_{i,j}$. The second matrix is further multiplied with the transpose of the second matrix, and the resulting matrix will be used for projecting feature vectors into a verification space, in which the actual comparison between the feature vectors for speaker verification will be
35 performed.

It should be noted, that the first matrix may be understood as a "small-variance" submatrix of the first principal component matrix and that the second matrix may be understood as a "large-variance" submatrix of the second principal component matrix.

Furthermore, it is a second object of the present invention to provide a method for speaker verification using a verification projection matrix formed in the above described way, with improved speaker verification results especially under noisy conditions.

To achieve the second object, according to a second aspect of the invention, a method for speaker verification using a verification projection matrix formed according to the method according to the first aspect of the invention is provided, wherein a speech signal of a requesting speaker is detected. The detected speech signal is time-aligned with the pre-stored reference speech signals of the true speaker. In a further step, feature vectors for the detected speech signal are generated from the time-aligned speech signal. Next, the feature vector of the i^{th} reference-aligned frame of the detected speech signal is projected into a projection space using the verification projection matrix and a distance of the projected feature vector in the projection space from a pre-stored centroid feature vector of the true speaker, is computed. The requesting speaker is accepted as the true speaker, if the distance satisfies a predefined criterion, e.g. if the distance is smaller than a pre-determined threshold.

With the present invention, it is now possible to increase the inter-speaker distance and at the same time having small intra-speaker distance.

It should be noted that the inter-speaker distance is the distance (in the verification space) between the feature

vectors of the true speaker and the feature vectors of the impostors.

Moreover, the intra-speaker distance is the distance (in the verification space) between different feature vectors of the same true speaker.

Thus, a feature vector from a true speaker differs more from a feature vector from an impostor, thereby increasing the security of distinguishing between the feature vectors and thus increasing the security of the speaker verification method compared with the methods according to the state of the art. This leads to still very good verification results even under noisy conditions.

According to a third aspect of the invention, a device for generating a verification projection matrix used for verification of an unknown speaker is provided, comprising a processor, which is formed such that the steps of the method according to the first aspect of the invention mentioned above may be performed.

According to a fourth aspect of the invention, a speaker verification system is provided comprising the device according to the third aspect of the invention, with

- a detecting unit for detecting a speech signal of a requesting speaker,
- a time-alignment unit for time-alignment of the detected speech signal with the pre-stored reference speech signals of the true speaker,
- a unit for forming feature vectors for the detected speech signal,
- a projection unit for projecting the feature vector of the i^{th} reference-aligned frame of the detected speech signal into a projection space using the verification projection matrix,

- a distance determination unit for computing a distance of the projected feature vector in the projection space from a pre-stored centroid feature vector of the true speaker,
- a verification unit for determining, whether the requesting speaker is accepted as the true speaker.

According to further aspects of the invention, a computer readable medium with a program recorded thereon and a computer program element are provided, where the program element makes the computer execute a procedure comprising the steps mentioned above.

The embodiments described in the following apply to the method, the device for generating a verification projection matrix, the speaker verification system, the computer readable medium, as well as to the computer program element.

There may be selected a plurality of first matrices from the first principal component matrix, and for each selected first matrix the following steps may be executed, thereby generating a second principal component matrix and a second matrix of the second principal component matrix for each selected first matrix:

- Projecting the impostor feature vectors of the i^{th} reference-aligned frame using the respective first matrix into a first space, thereby generating projected impostor feature vectors,
- performing a principal component analysis on the projected impostor feature vectors of the i^{th} reference-aligned frame, thereby generating a second set of eigenvectors, which form a respective second principal component matrix, and a second set of eigenvalues $\lambda'_{i,j}$, and
- selecting a second matrix from the second principal component matrix, wherein the second matrix comprises at least a part of the second set of eigenvectors associated with large eigenvalues.

Furthermore, the second matrix that yields the minimum equal error rate of verification results based on training data is selected as the verification projection matrix.

- 5 In this embodiment of the invention, the first matrices may have different dimensions.

By this embodiment, the verification results using the selected verification projection matrix are improved, as
10 there is used a plurality of first matrices, from each of which a respective second matrix is generated, thereby receiving a verification projection matrix with further smaller ratio of intra-speaker distance to inter-speaker distance.

15 If a first matrix is composed of eigenvectors from the first set of eigenvectors with corresponding u smallest eigenvalues, the first matrix is a matrix with corresponding eigenvectors, along which the variance of the feature vectors
20 of the i^{th} reference-aligned frame of the true speaker is minimal, thereby achieving a first projection matrix which projects feature vectors in a first space, in which the distance between feature vectors of the i^{th} reference-aligned frame of the true speaker, is minimal with reference to the
25 number of corresponding eigenvectors used in the first matrix. Thus, the intra-speaker distance for the i^{th} reference-aligned frame is minimal. Since the first space is specifically determined based on the data of the true speaker, the spread of the feature vectors from impostors
30 within the first space would not be identically small. This will provide further improved verification results.

If a second matrix is composed of eigenvectors from the second set of eigenvectors with corresponding p largest
35 eigenvalues, the second matrix is a matrix with corresponding eigenvectors, along which the variance of the distance between the feature vectors of the i^{th} reference-aligned

frame of the impostors and those of the true speaker is maximal, thereby achieving a second projection matrix which projects feature vectors in a second space, in which the distance between feature vectors of the i^{th} reference-aligned frame of the true speaker and feature vectors of the i^{th} reference-aligned frame of impostors is maximal with reference to the number of corresponding eigenvectors used in the second matrix. Thus, the inter-speaker distance for the i^{th} reference-aligned frame is large. Since the second space is specifically determined based on the data of the impostors the spread of the feature vectors from impostors within the second space and thus in the projection space, in which the feature vectors actually are compared in the verification mode, would further increase. This will provide further improved verification results.

According to a further embodiment of the invention, a verification projection matrix is generated for each reference-aligned frame of the whole speech signal, thereby generating r verification projection matrices. Using r verification projection matrices, the verification results improve further.

By this embodiment of the present invention, the verification result is further improved, as each true speaker has its own customized verification projection matrix for each reference-aligned frame.

The invention may be implemented in a programmable computer device as well as with a special electronic circuit.

BRIEF DESCRIPTION OF THE DRAWINGS

FIGURE 1 is a block diagram of a speaker verification system according to a preferred embodiment of the invention;

FIGURE 2A, 2B is a flow diagram of a speaker verification method including a training mode (Figure 2A) and a verification mode (Figure 2B) according to a preferred embodiment of the invention;

FIGURE 3 is a flow diagram of a method for generating verification projection matrices according to a preferred embodiment of the invention;

FIGURE 4A, 4B, 4C is a flow diagram of a method for generating a verification projection matrix for the i^{th} reference-aligned frame according to a preferred embodiment of the invention;

FIGURE 5A, 5B is a flow diagram of a method for generating a second matrix for a first matrix according to a preferred embodiment of the invention; and

FIGURE 6A, 6B is a flow diagram of a method for verification of a requesting speaker according to a preferred embodiment of the invention.

DETAILED DESCRIPTION OF A PREFERRED EMBODIMENT OF THE INVENTION

A preferred embodiment of the invention is now described with reference to the accompanying drawings.

1. Speaker verification system 100 (Overview)

As shown in **Fig.1** a speaker verification system 100 comprises:

- a detecting unit 101,
- a time-alignment unit 102,
- a feature vector estimation unit 103,

11

- a memory 104,
- a distance determination unit 105,
- a verification unit 106, and
- an output unit 107,

5 which are connected with each other via a computer bus 108.

The functionality of each unit will be described later in the context of the description of the method.

10 Each unit is formed in such a manner that it is able to perform the respective necessary steps.

The time-alignment unit 102, the feature vector estimation unit 103, the distance determination unit 105, and the
15 verification unit 106 may altogether or in part be implemented by a microprocessor, e.g. by a central processing unit of a personal computer.

2. Pre-Processing

20 For speaker verification, a requesting speaker 109 speaks a speech signal 110 into a microphone 111. The microphone 111 is connected to the speaker verification system 100, namely to the detecting unit 101, via a cable 112. Instead of the
25 cable 112, any other connection means may be used, for example an infrared interface. The detecting unit 101 is a sound card according to the preferred embodiment of the invention.

30 The input speech signal 110 is digitized by the detecting unit 101 at a sampling rate of 8 kHz. Subsequently, the speech signal is pre-emphasized by a first order digital filter having a transfer function $(1 - 0.99z^{-1})$. The resulting
35 filtered speech signal is then divided into overlapping frames with a frame size of 32 milliseconds (ms) and a two-third frame overlap.

Each of the generated frames is multiplied with a Hamming window function and is transformed into a frequency space using a Fast Fourier Transformation (FFT). Eventually, in the feature vector estimation unit 103, one feature vector is
5 generated for each frame consisting of n Mel-Frequency Cepstral Coefficients (MFCCs), each feature vector having dimension n , for each detected speech signal.

This kind of pre-processing is performed as well for detected
10 speech signals of a true speaker or an impostor in the training mode as for detected speech signals of a requesting speaker in the verification mode.

3. Method Overview

15 As shown in **Fig.2A**, in a first step, the entire method is started (step 201).

In a further step, the speaker verification system 100 is
20 trained for the speaker of interest in a training mode (step 202), in which step the necessary verification projection matrix/matrices, which will be described later in detail, is/are generated.

25 A first speaker c of interest speaks the predetermined speech signal 110 into the speaker verification system 100. The speech signal 110 according to the embodiment is the phrase "Hello". The first speaker's speech signal 110 is pre-processed and a feature vector with n MFCCs is computed for
30 each frame. The first speaker c repeatedly speaks the phrase "Hello" into the speaker verification system 100, thereby generating m speech signals 110, for which a feature vector with n MFCCs is computed for each frame of each speech signal.

Thus, there are provided m speech signals of the speaker of interest c . All feature vectors of the m speech signals are stored in the memory 104.

Furthermore, there are provided s speech signals spoken into the speaker verification system 100 from f impostors. Each speech signal is preprocessed and a feature vector with n MFCCs is computed for each frame. All impostor feature vectors are stored in the memory 104.

Note that a reference speech signal will be determined later on (see second paragraph of section 4 of the detailed description). Moreover, each speech signal will be time-aligned against the reference speech signal later on (see third paragraph of section 4 of the detailed description). Now let r be the total number of frames of the reference speech signal.

Moreover, $y_{i,j}$ is denoted as a frame of feature vector of the j^{th} speech signal of the true speaker (where $j = 1, \dots, m$), which frame is time-aligned (mapped) with the i^{th} frame of the reference speech signal in the time-alignment unit (where $i = 1, \dots, r$). This alignment is done for every frame of feature vector of the j^{th} speech signal of the true speaker, meaning that with every frame ($i = 1, \dots, r$) of the reference speech signal a frame of feature vector of the j^{th} speech signal of the true speaker is aligned, thereby producing r reference-aligned frames of feature vector of the j^{th} speech signal of the true speaker. In this context it should further be noted that the number of frames of feature vector of the speech signal of the true speaker may differ from the number r of frames of the reference speech signal. In this latter case, if the number of frames of the speech signal of the true speaker is higher than that of the reference speech signal, one or more frame(s) of the speech signal of the true speaker are dropped (i.e. are not aligned to any of the frames

of the reference speech signal), and, if the number of frames of the speech signal of the true speaker is lower than that of the reference speech signal, one or more frame(s) of the speech signal of the true speaker is/are each aligned to more than one frames of the reference speech signal.

For ease of description, it is referred to the index i as the i^{th} reference-aligned frame of the speech signal of the true speaker.

Moreover, let $\underline{x}_{i,k}$ denote the i^{th} ($i = 1, \dots, r$) reference-aligned frame of the k^{th} ($k = 1, \dots, s$) speech signal associated with impostors.

Using the stored feature vectors $\underline{y}_{i,j}$ and impostor feature vectors $\underline{x}_{i,k}$, the speaker verification system 100 is trained in the training mode (step 202) for the first speaker of interest c .

In a further step (step 203), it is determined, whether the speaker verification system 100 is to be trained for a further speaker of interest ($c + 1$).

If a further speaker of interest should be taken into consideration, the index c , denoting the respective speaker of interest, is increased by 1 (step 204) and the training is performed for the speaker of interest ($c + 1$).

If no further speaker of interest should be taken into consideration, the training mode (step 202) is accomplished (block 205) and the speaker verification system 100 is ready for a verification mode.

As shown in **Fig.2B**, in the verification mode, for a requesting speaker the identity is determined, that is the

15

requesting speaker is verified by the speaker verification system 100 (step 206).

5 As a result of the verification, the speaker verification system 100, namely the output unit 107 presents the result 113 to the requesting speaker, either the acceptance or the denial of the requesting speaker (cf. **Fig.1**).

10 In a last step (step 207), the program for speaker verification stops.

4. Detailed description of the training mode (step 202)

15 Referring now to **Fig.3**, the training mode (step 202) will be described in detail.

20 In a first step, according to the method described in (J.G. Wilpon and L.R. Rabiner, A Modified K-Means Clustering Algorithm for Use in Isolated Word Recognition, IEEE Transactions on Acoustics Speech and Signal Processing, Volume 33, Pages 587 - 594, 1985), a reference speech signal is determined.

25 In a further step, all stored feature vectors $\underline{y}_{i,j}$ and impostor feature vectors $\underline{x}_{i,k}$ are time-aligned ($\underline{y}_{i,j}, \underline{x}_{i,k} \in \mathbb{R}^{n \times 1}$) with the frames of the reference speech signal of the true speaker in the time-alignment unit 102.

30 After having started to generate verification projection matrices (step 301, cf. **Fig.3**), a verification projection matrix is generated for the i^{th} reference-aligned frame (step 302).

16

The computed verification projection matrix for the i^{th} reference-aligned frame for the speaker c of interest, is stored in the memory 104 (step 303).

- 5 In a further step (step 304), it is determined, whether a further verification projection matrix for a further frame is to be generated.

10 If a further verification projection matrix is to be generated, the frame index i is increased by 1 (step 305) and the verification projection matrix is generated for the reference-aligned frame $i + 1$.

15 If there has been generated a verification projection matrix for each reference-aligned frame of the speech signals of the speaker of interest, the program returns to step 203 as shown in **Fig.2A** (step 306).

20 **Fig.4A, Fig.4B** and **Fig.4C** show a flow diagram of the steps being executed for generating a verification projection matrix for a frame i .

25 Since all speech signals are time-aligned as described above, it is possible to perform an analysis for one frame of all speech signals at a time.

After a starting step (step 401), for the i^{th} reference-aligned frame, a centroid feature vector $\bar{y}_i$ of the feature vectors $y_{i,1}, \dots, y_{i,m}$ is determined according to the following
30 formula (step 402) and is stored in the memory 104:

$$\bar{y}_i = \frac{1}{m} \sum_{j=1}^m y_{i,j}. \quad (1)$$

17

In a further step (step 403), a difference feature vector $\hat{\underline{y}}_{i,j}$ is computed for the i^{th} reference-aligned frame of the speech signal j by subtracting the centroid feature vector $\bar{\underline{y}}_i$ from the feature vector $\underline{y}_{i,j}$.

5

Thus, the difference feature vector $\hat{\underline{y}}_{i,j}$ is determined according to the following formula:

$$10 \quad \hat{\underline{y}}_{i,j} = \underline{y}_{i,j} - \bar{\underline{y}}_i. \quad (2)$$

The determined difference feature vector $\hat{\underline{y}}_{i,j}$ is stored in the memory 104 (step 404) and it is determined (step 405), whether all difference feature vectors $\hat{\underline{y}}_{i,j}$ for all feature
15 vectors $\underline{y}_{i,j}$ have been computed.

If there is still a further speech signal, the difference feature vector $\hat{\underline{y}}_{i,j+1}$ for the next speech signal is computed (speech signal index j is increased by 1 (step 406)).

20

In a further step (step 407), a first covariance matrix $\underline{Y}_i$ for the i^{th} reference-aligned frame of the feature vectors $\underline{y}_{i,j}$ is computed according to the following formula:

$$25 \quad \underline{Y}_i = \frac{1}{m} \cdot \hat{\underline{Y}}_i \cdot \hat{\underline{Y}}_i^T, \quad (3)$$

wherein

$$\hat{\underline{Y}}_i = [\hat{\underline{y}}_{i,1}, \dots, \hat{\underline{y}}_{i,m}] \quad (4)$$

30

and

$$\underline{Y}_i \in \mathfrak{R}^{n \times n} \quad (5)$$

and

$$\underline{\hat{Y}}_i \in \mathfrak{R}^{n \times m}. \quad (6)$$

5 By solving the eigenvalue problem, i.e. by solving the equation system,

$$(\underline{Y}_i - \lambda_{i,t} \cdot \underline{1}) \cdot \underline{v}_{i,t} = \underline{0}, \quad (7)$$

10 wherein

- $\lambda_{i,t}$ is the t^{th} eigenvalue of the equation system
($t = 1, \dots, n$),
- $\underline{1}$ is the identity matrix of dimension n , and
- 15 • $\underline{v}_{i,t}$ is the corresponding t^{th} eigenvector of the equation system,

thereby performing a first principal component analysis, the eigenvalues $\lambda_{i,t}$ and the corresponding eigenvectors $\underline{v}_{i,t}$ are determined (step 408) for the first covariance matrix $\underline{Y}_i$ for
20 the i^{th} reference-aligned frame.

Such eigenvectors $\underline{v}_{i,t}$ are denoted as the first set of eigenvectors and the corresponding eigenvalues $\lambda_{i,t}$ as the first set of eigenvalues.

25 The eigenvalue problem is solved by solving the characteristic equation, which is a polynomial of degree n :

$$\det(\underline{Y}_i - \lambda_{i,t} \cdot \underline{1}) = 0. \quad (8)$$

30 Since the first covariance matrix $\underline{Y}_i$ is non-negative definite, its eigenvectors are chosen to be orthonormal and the eigenvalues are non-negative.

35 Thus, according to the present embodiment, it is assumed that

$$\lambda_{i,1} \leq \lambda_{i,2} \leq \dots \leq \lambda_{i,t} \leq \dots \leq \lambda_{i,n}. \quad (9)$$

The variance of the projections of the feature vectors $\underline{y}_{i,j}$ onto the respective eigenvector $\underline{v}_{i,t}$ is thus quantified based on the corresponding eigenvalue $\lambda_{i,t}$, namely it is proportional to the eigenvalue $\lambda_{i,t}$.

Thus, it should be understood, that if a feature vector of a speaker c is projected into a first space defined by a matrix, which comprises some eigenvectors associated with small eigenvalues of the first principal component matrix, the spread of the projected feature vectors would be rather small, thus giving rise to a small intra-speaker distance. The first principal component matrix is defined as the matrix consisting of all the determined eigenvectors $\underline{v}_{i,t}$, and is determined in a further step (step 409).

Since the first space for each speaker is specifically determined based on the feature vectors of the i^{th} reference-aligned frame of the m speech signals of the speaker of interest c , the spread of projected feature vectors from an impostor would not be identically small. Therefore, good verification results could be expected.

After having initialized k to the value 1 (step 410), in a further step (step 411), a first matrix $\underline{A}_{i,u}$ is selected from the first principal component matrix, wherein

$$\underline{A}_{i,u} = [\underline{v}_{i,1}, \underline{v}_{i,2}, \dots, \underline{v}_{i,u}], \quad (10)$$

with

$$u = p + k - 1, \text{ and} \quad (11)$$

p being an arbitrarily predetermined parameter.

20

Thus, the first matrix $\underline{A}_{i,u}$ comprises the u eigenvectors associated with the u smallest eigenvalues.

For the selected first matrix $\underline{A}_{i,u}$, a first projection
 5 matrix $\underline{P}_{i,u}$ is computed according to the following formula
 (step 412):

$$\underline{P}_{i,u} = \underline{A}_{i,u} \cdot \underline{A}_{i,u}^T. \quad (12)$$

10 In a further step (step 413), the first projection matrix
 $\underline{P}_{i,u}$ is stored in the memory 104.

In a sequence of further steps, in **Fig. 4B** symbolized as step
 414, a second matrix is computed for the first projection
 15 matrix $\underline{P}_{i,u}$.

The method for computing the second matrix is shown in **Fig. 5A**
 and **Fig. 5B**.

20 After having started the method for computing the second
 matrix (step 501), a difference impostor feature vector $\hat{x}_{i,k}$
 for the i^{th} reference-aligned frame of the k^{th} impostor
 speech signal is computed (step 502) by subtracting the
 centroid feature vector of the true speaker $\bar{y}_i$ from the
 25 corresponding impostor feature vector $x_{i,k}$ according to the
 following formula:

$$\hat{x}_{i,k} = x_{i,k} - \bar{y}_i. \quad (13)$$

30 The difference impostor feature vector $\hat{x}_{i,k}$ is stored in the
 memory 104 (step 503).

Furthermore, it is determined (step 504), whether all
 difference impostor feature vectors $\hat{x}_{i,k}$ for all impostor
 35 feature vectors $x_{i,k}$ have been computed.

If there is still a further impostor speech signal, then the difference impostor feature vector for the next impostor speech signal is computed (speech signal index k is increased by 1 (step 505)).

5

In a further step (step 506), each difference impostor feature vector $\hat{x}_{i,k}$ is projected into a first space using the first projection matrix $P_{i,u}$ according to the following

formula, thereby determining s projected difference impostor
 10 feature vectors $z_{i,k,u}$:

$$z_{i,k,u} = P_{i,u} \cdot \hat{x}_{i,k}. \quad (14)$$

In a further step (step 507), a centroid projected difference
 15 impostor feature vector $\bar{z}_{i,u}$ is determined according to the following formula:

$$\bar{z}_{i,u} = \frac{1}{s} \sum_{k=1}^s z_{i,k,u}. \quad (15)$$

20 In a further step (step 508), a second covariance matrix $\underline{Z}_{i,u}$ for the i^{th} reference-aligned frame of the projected difference impostor feature vectors $z_{i,k,u}$ is computed according to the following formula:

$$25 \quad \underline{Z}_{i,u} = \frac{1}{s} \cdot \hat{\underline{Z}}_{i,u} \cdot \hat{\underline{Z}}_{i,u}^T, \quad (16)$$

wherein

$$\hat{\underline{Z}}_{i,u} = [\hat{z}_{i,1,u}, \dots, \hat{z}_{i,s,u}], \quad (17)$$

30

and

$$\hat{z}_{i,k,u} = z_{i,k,u} - \bar{z}_{i,u}. \quad (18)$$

By solving the eigenvalue problem, i.e. by solving the equation system,

$$(\underline{Z}_{i,u} - \hat{\lambda}_{i,t,u} \cdot \underline{1}) \cdot \hat{\underline{v}}_{i,t,u} = \underline{0}, \quad (19)$$

5

wherein

- $\hat{\lambda}_{i,t,u}$ is the t^{th} eigenvalue of the equation system
($t = 1, \dots, n$),
- $\underline{1}$ is the identity matrix of dimension n , and
- 10 • $\hat{\underline{v}}_{i,t,u}$ is the corresponding t^{th} eigenvector of the equation system,

thereby performing a second principal component analysis, the eigenvalues $\hat{\lambda}_{i,t,u}$ and the corresponding eigenvectors $\hat{\underline{v}}_{i,t,u}$ are determined (step 509) for the second covariance matrix $\underline{Z}_{i,u}$ for the i^{th} reference-aligned frame.

15

Such eigenvectors $\hat{\underline{v}}_{i,t,u}$ are denoted as the second set of eigenvectors, and the corresponding eigenvalues $\hat{\lambda}_{i,t,u}$ are denoted as the second set of eigenvalues.

20

The eigenvalue problem is solved by solving the characteristic equation, which is a polynomial of degree n :

$$\det(\underline{Z}_{i,u} - \hat{\lambda}_{i,t,u} \cdot \underline{1}) = 0. \quad (20)$$

25

Since the second covariance matrix $\underline{Z}_{i,u}$ is non-negative definite, its eigenvectors $\hat{\underline{v}}_{i,t,u}$ is chosen to be orthonormal and the eigenvalues $\hat{\lambda}_{i,t,u}$ is non-negative.

30 Thus, according to the present embodiment of the invention, it is assumed that

$$\hat{\lambda}_{i,1,u} \leq \hat{\lambda}_{i,2,u} \leq \dots \leq \hat{\lambda}_{i,t,u} \leq \dots \leq \hat{\lambda}_{i,n,u}. \quad (21)$$

35 From the relation

$$\hat{\lambda}_{i,t,u} = \sum_{k=1}^S \left(\hat{\underline{v}}_{i,t,u}^T \cdot \underline{z}_{i,k,u} \right)^2, \quad (22)$$

it should be understood, that the second set of eigenvectors
 5 $\hat{\underline{v}}_{i,t,u}$ with large eigenvalues $\hat{\lambda}_{i,t,u}$ constitute a subspace
 with large inter-speaker distance.

Moreover, such a subspace has small intra-speaker distance,
 since it is a subspace of the first space which has small
 10 intra-speaker distance. Such a subspace is what is desired in
 order to get good verification results.

The second principal component matrix is defined as the
 matrix consisting of the determined second eigenvectors
 15 $\hat{\underline{v}}_{i,t,u}$ and is determined in a further step (step 510).

From the second principal component matrix, a second matrix
 consisting of the p second eigenvectors associated with the
 corresponding p largest eigenvalues (of the second set of
 20 eigenvalues) is selected (step 511) and stored in the memory
 104 (step 512).

In a further step (step 513), the program returns to
 step 415. In a further step (step 415), using the second
 25 matrix, the equal error rate of the verification result based
 on the training data is computed.

In a further step (step 416) it is determined, whether a
 larger number of eigenvectors (from the first set of
 30 eigenvectors) should be selected for computing another first
 projection matrix.

If a larger number of eigenvectors is to be selected, the
 index k is increased by 1 (step 417). The program goes back

to step 411 and another second matrix is generated in the same way as described above.

Thus, it should be understood, that there may be selected $n-p+1$ different first projection matrices at maximum, as there may be selected a first projection matrix with only p first eigenvectors associated with the lowest p eigenvalues up to a first projection matrix with all first eigenvectors.

Under these eventually generated $n-p+1$ second matrices, in a further step, the second matrix with the minimum equal error rate (denoted as $\hat{A}_{i,\hat{u}}$) is selected (step 418).

In a further step (step 419) a second projection matrix $\hat{P}_{i,\hat{u}}$ is computed as verification projection matrix for the selected second matrix $\hat{A}_{i,\hat{u}}$

$$\hat{P}_{i,\hat{u}} = \hat{A}_{i,\hat{u}} \cdot \hat{A}_{i,\hat{u}}^T. \quad (23)$$

In a further step of the method for generating the verification projection matrix for the i^{th} reference-aligned frame (step 420), the program returns to step 303.

In a further step, the second projection matrix $\hat{P}_{i,\hat{u}}$ is stored in the memory 104 (step 303).

Now, the training of the speaker verification system 100 is complete.

Before the beginning of the verification, in the memory 104, there is stored for each true speaker, and for each frame of the speech signal of the respective speaker a determined verification projection matrix.

5. Detailed description of the verification mode (step 206)

Referring now to **Fig. 6A** and **Fig. 6B**, the verification mode (step 206) will be described in detail.

- 5 After having started the program for verification of a requesting speaker (step 601), the requesting speaker 109 is prompted to speak the predetermined speech signal 110 into the microphone 111.
- 10 As described under section 1 (Speaker verification system 100 (Overview)) above, the speech signal of the requesting speaker is detected (step 602).

15 In a further step, the frames of the speech signal spoken in are time-aligned with the frames of the reference speech signal of the true speaker, thereby generating r reference-aligned frames and corresponding feature vectors for the detected speech signal, where r is the total number of frames of the reference speech signal (step 603).

- 20 Furthermore, the feature vector of each reference-aligned frame i is respectively projected into a space using the verification projection matrix for the i^{th} reference-aligned frame of the true speaker (step 604), thus generating a
- 25 projected feature vector $\hat{w}_i$ according to the following formula:

$$\hat{w}_i = \hat{p}_{i,\hat{u}} \cdot w_i. \quad (24)$$

- 30 In a further step (step 605), by the distance determination unit 105, a frame distance D_i of the projected feature vector $\hat{w}_i$ from the centroid feature vector of the true speaker (of the i^{th} reference-aligned frame) is determined according to the following regulation:

$$D_i = \|\hat{p}_{i,\hat{u}} \cdot w_i - \hat{p}_{i,\hat{u}} \cdot \bar{y}_i\|. \quad (25)$$

In a further step (step 606), the determined frame distance D_i is stored in the memory 104 and it is checked whether a further frame distance D_{i+1} is to be determined (step 607).

5

If a further frame distance is to be determined, the frame index i is increased by 1 (step 608) and the next frame distance D_{i+1} is determined.

10 If all frame distances D_i are determined, a total distance D of all frame distances is determined (step 609) according to the following formula:

$$D = \sum_{i=1}^r D_i. \quad (26)$$

15

The total distance D is stored in the memory 104 (step 610).

In a further step (step 611) it is determined (by the verification unit 106), whether the total distance D is
20 smaller than a predetermined threshold.

If the total distance D is smaller than the threshold or equal to the threshold, the requesting speaker is accepted as the true speaker (step 612).

25

If the total distance is larger than the threshold, the verification of the requesting speaker is denied (step 613).

It has been shown in experiments, that very good verification
30 results were achieved even with the parameter p chosen to be equal to 1 and using 10 speech signals each from 15 persons (1 speaker of interest and 14 impostors) for training.

CLAIMS

What is claimed is:

- 5 1. A method for generating a verification projection matrix used for verification of an unknown speaker, wherein
- 10 a) a feature vector for each reference-aligned frame i ($i = 1, \dots, r$) of each speech signal out of a plurality of time-aligned speech signals j ($j = 1, \dots, m$) by the true speaker, is provided,
- 15 b) an impostor feature vector for each reference-aligned frame i ($i = 1, \dots, r$) of each speech signal out of a plurality of digitized time-aligned speech signals k ($k = 1, \dots, s$) by several impostors is provided,
- c) for all feature vectors of the i^{th} reference-aligned frame, the following steps are performed:
- 20 c1) Performing a principal component analysis on the feature vectors of the i^{th} reference-aligned frame of the true speaker, thereby generating a first set of eigenvectors, which form a first principal component matrix, and a first set of eigenvalues $\lambda_{i,j}$;
- 25 c2) Selecting a first matrix from the first principal component matrix, wherein the first matrix comprises at least a part of the eigenvectors from the first set of eigenvectors associated with small eigenvalues $\lambda_{i,j}$;
- 30 c3) Projecting the impostor feature vectors of the i^{th} reference-aligned frame using the first matrix into a first space, thereby generating projected impostor feature vectors;
- 35 c4) Performing a principal component analysis on the projected impostor feature vectors of the i^{th} reference-aligned frame, thereby generating a

second set of eigenvectors, which form a second principal component matrix, and a second set of eigenvalues $\lambda'_{i,j}$;

- 5 c5) Obtaining a second matrix as the verification projection matrix from the second principal component matrix, wherein the second matrix comprises at least a part of the second set of eigenvectors associated with large eigenvalues $\lambda'_{i,j}$.

10 2. A method for generating a verification projection matrix used for verification of an unknown speaker according to claim 1, wherein

- 15 • a plurality of first matrices is selected from the first principal component matrix in step c2),
• the steps c3) and c4) are performed for each selected first matrix, thereby generating a plurality of second principal component matrices,
20 • for each second principal component matrix, a second matrix is formed comprising at least a part of the second eigenvectors associated with large eigenvalues,
• in step c5), among all the second matrices the one is selected that yields the minimum equal error rate of
25 verification results based on training data.

3. A method for generating a verification projection matrix used for verification of an unknown speaker according to claim 2,

30 wherein the first matrices have different dimensions.

4. A method for generating a verification projection matrix used for verification of an unknown speaker according to claim 1,

35 wherein the first matrix is a matrix composed of the u eigenvectors from the first set of eigenvectors with corresponding u smallest eigenvalues.

5. A method for generating a verification projection matrix used for verification of an unknown speaker according to claim 1,

5 wherein the second matrix is a matrix composed of the p eigenvectors from the second set of eigenvectors with corresponding p largest eigenvalues.

10 6. A method for generating a verification projection matrix used for verification of an unknown speaker according to claim 1,

wherein a verification projection matrix is generated for each reference-aligned frame of the whole speech signal, thereby generating r verification projection matrices.

15 7. A method for generating a verification projection matrix used for verification of an unknown speaker according to claim 1,

- 20 • wherein a feature vector for each reference-aligned frame i ($i = 1, \dots, r$) of each speech signal out of a plurality of time-aligned speech signals j ($j = 1, \dots, m$) by a plurality of true speakers, is provided, and
- wherein the method is performed for each speaker out of the plurality of true speakers, thereby generating as many
25 projection matrices as there are true speaker for each reference-aligned frame i.

8. A method for generating a verification projection matrix used for verification of an unknown speaker according to
30 claim 7,

wherein a verification projection matrix is generated for each reference-aligned frame of the whole speech signal of each true speaker, thereby generating r verification projection matrices for each true speaker.

35

9. A method for speaker verification using a verification projection matrix formed according to the method according to claim 1,

wherein

- 5 • a speech signal of a requesting speaker is detected,
- feature vectors for the detected speech signal are formed,
- the detected speech signal is time-aligned with the pre-stored reference speech signals of the true speaker,
- the feature vector of the i^{th} reference-aligned frame of
- 10 the detected speech signal is projected into a projection space using the verification projection matrix,
- a distance of the projected feature vector in the projection space from a pre-stored centroid feature vector of the true speaker, is computed,
- 15 • the requesting speaker is accepted as the true speaker, if the distance satisfies a predefined criterion.

10. A method for speaker verification according to claim 9, wherein the distance is the Euclidian distance of the

20 projected feature vector from a pre-stored centroid feature vector of the true speaker in the projection space.

11. A method for speaker verification according to claim 9, wherein

- 25 • a verification projection matrix is generated for each reference-aligned frame of the whole speech signal, thereby generating r verification projection matrices,
- the steps of the method according to claim 9 are performed for each reference-aligned frame of the detected time-
- 30 aligned speech signal, and
- the requesting speaker is accepted as the true speaker, if the total distances of all reference-aligned frames satisfy a predefined criterion.

35 12. A device for generating a verification projection matrix used for verification of an unknown speaker, wherein a processor is provided, which may perform the following steps:

31

- a) a feature vector for each reference-aligned frame i ($i = 1, \dots, r$) of each speech signal out of a plurality of time-aligned speech signals j ($j = 1, \dots, m$) by the true speaker, is provided,
- 5 b) an impostor feature vector for each reference-aligned frame i ($i = 1, \dots, r$) of each speech signal out of a plurality of digitized time-aligned speech signals k ($k = 1, \dots, s$) by several impostors, is provided,
- c) for all feature vectors of the i^{th} reference-aligned frame, the following steps are performed:
- 10 c1) Performing a principal component analysis on the feature vectors of the i^{th} reference-aligned frame of the true speaker, thereby generating a first set of eigenvectors, which form a first principal component matrix, and a first set of eigenvalues
- 15 $\lambda_{i,j}$,
- c2) Selecting a first matrix from the first principal component matrix, wherein the first matrix comprises at least a part of the eigenvectors from the first set of eigenvectors associated with small
- 20 eigenvalues $\lambda_{i,j}$,
- c3) Projecting the impostor feature vectors of the i^{th} reference-aligned frame using the first matrix into a first space, thereby generating projected impostor feature vectors,
- 25 c4) Performing a principal component analysis on the projected impostor feature vectors of the i^{th} reference-aligned frame, thereby generating a second set of eigenvectors, which form a second principal component matrix, and a second set of eigenvalues
- 30 $\lambda'_{i,j}$,
- c5) Obtaining a second matrix as the verification projection matrix from the second principal component matrix, wherein the second matrix
- 35 comprises at least a part of the second set of

eigenvectors associated with large eigenvalues
 $\lambda'_{i,j}$.

13. A speaker verification system comprising a device
5 according to claim 12, comprising
- a detecting unit for detecting a speech signal of a requesting speaker,
 - a unit for forming feature vectors for the detected speech signal,
 - 10 • a time-alignment unit for time-alignment of the detected speech signal with the pre-stored reference speech signals of the true speaker,
 - a projection unit for projecting the feature vector of the i^{th} reference-aligned frame of the detected speech signal
15 into a projection space using the verification projection matrix,
 - a distance determination unit for computing a distance of the projected feature vector in the projection space from a pre-stored centroid feature vector of the true speaker,
 - 20 • a verification unit for determining, whether the requesting speaker is accepted as the true speaker.

14. A computer readable storage medium, having a program
recorded thereon, where the program makes the computer
25 execute a procedure comprising the following steps for generating a verification projection matrix used for verification of an unknown speaker, wherein
- a) a feature vector for each reference-aligned frame i ($i = 1, \dots, r$) of each speech signal out of a plurality of
30 time-aligned speech signals j ($j = 1, \dots, m$) by the true speaker is provided,
 - b) an impostor feature vector for each reference-aligned frame i ($i = 1, \dots, r$) of each speech signal out of a plurality of digitized time-aligned speech signals k ($k = 1, \dots, s$) by several impostors is provided,
35
 - c) for all feature vectors of the i^{th} reference-aligned frame, the following steps are performed:

33

- 5 c1) Performing a principal component analysis on the feature vectors of the i^{th} reference-aligned frame of the true speaker, thereby generating a first set of eigenvectors, which form a first principal component matrix, and a first set of eigenvalues $\lambda_{i,j}$,
- 10 c2) Selecting a first matrix from the first principal component matrix, wherein the first matrix comprises at least a part of the eigenvectors from the first set of eigenvectors associated with small eigenvalues $\lambda_{i,j}$,
- 15 c3) Projecting the impostor feature vectors of the i^{th} reference-aligned frame using the first matrix into a first space, thereby generating projected impostor feature vectors,
- 20 c4) Performing a principal component analysis on the projected impostor feature vectors of the i^{th} reference-aligned frame, thereby generating a second set of eigenvectors, which form a second principal component matrix, and a second set of eigenvalues $\lambda'_{i,j}$,
- 25 c5) Obtaining a second matrix as the verification projection matrix from the second principal component matrix, wherein the second matrix comprises at least a part of the second set of eigenvectors associated with large eigenvalues $\lambda'_{i,j}$.

15. A computer readable storage medium, having a program
30 recorded thereon, where the program makes the computer execute a procedure comprising the following steps for speaker verification using a verification projection matrix formed according to the method according to claim 1, wherein

- 35
- a speech signal of a requesting speaker is detected,
 - feature vectors for the detected speech signal are formed,

- the detected speech signal is time-aligned with the pre-stored reference speech signals of the true speaker,
- the feature vector of the i^{th} reference-aligned frame of the detected speech signal is projected into a projection space using the verification projection matrix,
- a distance of the projected feature vector in the projection space from a pre-stored centroid feature vector of the true speaker, is computed,
- the requesting speaker is accepted as the true speaker, if the distance satisfies a predefined criterion.

16. A computer program element which makes the computer execute a procedure comprising the following steps for generating a verification projection matrix used for verification of an unknown speaker, wherein

- a) a feature vector for each reference-aligned frame i ($i = 1, \dots, r$) of each speech signal out of a plurality of time-aligned speech signals j ($j = 1, \dots, m$) by the true speaker is provided,
- b) an impostor feature vector for each reference-aligned frame i ($i = 1, \dots, r$) of each speech signal out of a plurality of digitized time-aligned speech signals k ($k = 1, \dots, s$) by several impostors is provided,
- c) for all feature vectors of the i^{th} reference-aligned frame, the following steps are performed:
 - c1) Performing a principal component analysis on the feature vectors of the i^{th} reference-aligned frame of the true speaker, thereby generating a first set of eigenvectors, which form a first principal component matrix, and a first set of eigenvalues $\lambda_{i,j}$,
 - c2) Selecting a first matrix from the first principal component matrix, wherein the first matrix comprises at least a part of the eigenvectors from the first set of eigenvectors associated with small eigenvalues $\lambda_{i,j}$,

- c3) Projecting the impostor feature vectors of the i^{th} reference-aligned frame using the first matrix into a first space, thereby generating projected impostor feature vectors,
- 5 c4) Performing a principal component analysis on the projected impostor feature vectors of the i^{th} reference-aligned frame, thereby generating a second set of eigenvectors, which form a second principal component matrix, and a second set of
- 10 eigenvalues $\lambda'_{i,j}$,
- c5) Obtaining a second matrix as the verification projection matrix from the second principal component matrix, wherein the second matrix comprises at least a part of the second set of
- 15 eigenvectors associated with large eigenvalues $\lambda'_{i,j}$.

17. A computer program element which makes the computer execute a procedure comprising the following steps for
- 20 speaker verification using a verification projection matrix formed according to the method according to claim 1, wherein
- a speech signal of a requesting speaker is detected,
 - feature vectors for the detected speech signal are formed,
 - 25 • the detected speech signal is time-aligned with the pre-stored reference speech signals of the true speaker,
 - the feature vector of the i^{th} reference-aligned frame of the detected speech signal is projected into a projection space using the verification projection matrix,
 - 30 • a distance of the projected feature vector in the projection space from a pre-stored centroid feature vector of the true speaker, is computed,
 - the requesting speaker is accepted as the true speaker, if the distance satisfies a predefined criterion.