

A Speaker Recognition Method Based on Trajectories in Feature Spaces of Voice Samples

Technical Field

The invention relates to the field of biometrics, in particular to a speaker recognition method based on trajectories in feature spaces of voice samples.

Technical Background

With the development of artificial intelligence technology, audio perception has become a hotspot in audio processing technology research, where audio classification or audio recognition is the core issue of audio perception. In engineering applications, audio classification is represented by speaker recognition, audio event recognition, and audio event detection etc. Speaker recognition technology is a kind of identity verification technology - a type of biometric technology. Biometric technology is a technology that uses biological characteristics to automatically identify an individual's identity, comprising fingerprint recognition, iris recognition, gene recognition, face recognition etc. Compared with other authentication technologies, speaker identification is more convenient and natural, and has relatively low user intrusion. Speaker recognition uses voice signals for identity recognition, which has the advantages of natural human-computer interaction, easy extraction of voice signals, and realising remote recognition etc.

The existing speaker recognition system comprises two stages: training stage and recognition stage. In the training stage, the system uses the collected speaker voice to build a model of the speaker; in the recognition stage, the system matches the input voice with the speaker model to make a decision. The speaker recognition system needs to extract features that can reflect the personality of the speaker from the voice signal, and establish an accurate model to distinguish the difference between the speaker and other speakers. There are two main categories of commonly used audio classification techniques. One is generative statistical models, such as Gaussian Mixture Model GMM and Hidden Markov Model HMM, and the other is methods based on deep neural networks, such as DNN, RNN or LSTM etc. Regardless of the technology, a large number of labelled training samples are required. In order to achieve better recognition performance, the deep neural network methods require higher sample size. The methods based on GMM or HMM do not give special consideration to the distinguishing information between different audio categories, nor do they consider the sharing of different types of sample data, such as: the methods mentioned in MIT Professor Reynold et al.'s paper "Speaker Verification Using Adapted Gaussian Mixture Models" (Digital Signal Processing 10 (2000), 19-41.), which have high computational complexity; with the support of large samples, deep neural network methods have shown good performance, such as using neural networks to extract features and train speech in Google's paper "End-to-End Text-Dependent Speaker Verification" (2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2016, Pages: 5115-5119), but neural networks training requires a lot of labelled voices, and the acquisition cost of a large number of samples is very high, and the deep neural network method lacks interpretability, which is comparable to a black box.

Existing speaker recognition technologies often have high computational complexity and require a large amount of labelled speaker voice data to train the model, and collecting a large amount of labelled voice data requires a huge workload. Therefore, it is necessary to find a more convenient and effective speaker recognition method and system.

Summary of the invention

The purpose of the present invention is to provide a speaker recognition method based on trajectories in feature spaces of voice samples to address the shortcomings of the prior art. The voice feature space is independent of the speaker, text and language, the construction of the voice feature space can use any qualified voice data, which results in the sharing of voice data; and the speaker's voice trajectory can be constructed even with one sample, so there is no need of a large amount of labelled voice data, which overcomes the prior art deficiency which needs to collect a large amount of labelled voice data.

The purpose of the present invention can be achieved through the following technical solutions:

A speaker recognition method based on trajectories in feature spaces of voice samples, wherein a voice sample can be regarded as a movement in a voice feature space, and has characteristics of an active space and a trajectory in a space, the method comprises the following steps:

step 1) constructing the voice feature space Ω : clustering unlabelled voice samples in the feature space using a clustering method, generating, based on subcategory data obtained from the clustering, a certain expression of the subcategory data as an expression of the voice feature space $\Omega = \{g_k, k = 1, 2, \dots, K\}$;

step 2) constructing speaker knowledge: using clean voice samples labelled with speaker attributes to obtain corresponding distribution information and movement trajectory information in the voice feature space Ω ;

step 3) speaker recognition: for a voice sample to be recognized, first obtaining a distribution expression and a trajectory in the voice feature space of the sample, then using the distribution information in the voice feature space of a speaker to calculate a difference between a sample distribution and a prior distribution, and a cumulative local distribution difference along the trajectory, as a basis for a speaker identification and to perform a determination.

Further, in a process of constructing the voice feature space Ω of the step 1), any clean voice sample can be used, and there is no constraint on factors of speakers and languages.

Further, in the step 1), using K-means or other clustering methods to cluster the voice samples in the feature space, the expression of the voice feature space $\Omega = \{g_k, k = 1, 2, \dots, K\}$ can be a distribution function of class data (such as Gaussian distribution function), cluster center vector (centroid) or generative model (such as hidden Markov model or neural network), which are identifiers with positioning ability, namely, feature space identifiers, a class identifier size K used in the voice feature space determines a granularity of the expression of the voice feature space, the larger the K , the finer the expression of the voice feature space. On the other hand, the accuracy of space expression is related to the size of data. The richer the data, the more complete the space expression. At the same time, the more targeted the data to construct the voice feature space, the more accurate the space expression will be for a specific problem.

Further, in the step 2), using the clean voice samples labelled with speaker attributes to label the voice feature space, when using a Gaussian distribution $g_k(m_k, U_k)$ as a space identifier, the distribution information in the feature space of the speaker is obtained as follows:

i. calculating a position correlation between each feature f_t of the voice sample and the space identifier $g_k(m_k, U_k)$, which is defined as:

$$p_t^k = \frac{1}{\left[(f_t - m_k)^T U_k^{-1} (f_t - m_k) \right]^\alpha}$$

in the formula, the space identifier is represented by a multi-dimensional Gaussian distribution, m_k represents a mean vector of the k-th Gaussian distribution, U_k represents a variance matrix of the k-th multi-dimensional Gaussian distribution;

ii. calculating an expected value of the position correlation between a speaker sample set and the space identifier $g_k(m_k, U_k)$:

$$B^k = \sum_n \sum_t p_{n,t}^k$$

in the formula, $p_{n,t}^k$ represents a correlation between the t -th frame feature of the n -th sample and the space identifier $g_k(m_k, U_k)$;

iii. calculating the distribution in the feature space of the speaker as:

$$p_k = \frac{B^k}{\sum_k B^k}.$$

Further, in the step 2), a time sequence information of the movement trajectory of the speaker voice sample in the voice feature space Ω is expressed as a neighborhood sequence $\Psi_1 \Psi_2 \cdots \Psi_T$ of the voice sample feature, and a δ neighborhood of the voice sample feature f_t is $\Psi_t = \{g_k | d_{tk} < \delta\}$, where d_{tk} is the Mahalanobis distance between the voice sample feature and a distribution of the voice sample, that is $d_{tk} = \sqrt{(f_t - m_k)^T U_k^{-1} (f_t - m_k)}$.

Further, a decision threshold of the δ neighborhood $\Psi_t = \{g_k | d_{tk} < \delta\}$ of the voice sample feature f_t refers to characteristics of a normal distribution, where $2 < \delta < 3$ is selected.

Further, in the step 3), the speaker recognition process of the voice samples $f = \{f_1, f_2, \cdots, f_T\}$ comprises the following steps:

i. calculating the distribution $P = (p_1, p_2, \cdots, p_K)$ of voice sample $f = \{f_1, f_2, \cdots, f_T\}$ in the voice

feature space Ω , where $p_k = \frac{\sum_t p_t^k}{\sum_k \sum_t p_t^k}$;

ii. determining the movement trajectory $\Psi_1 \Psi_2 \cdots \Psi_T$ of the voice sample $f = \{f_1, f_2, \cdots, f_T\}$ in the voice feature space Ω , $\Psi_t = \{g_k | d_{tk} < \delta\}$;

iii. calculating a distance $d_s = \|P - P_s\|^\beta = \sum_k (p_k - p_k^s)^\beta$ between the distribution $P = (p_1, p_2, \cdots, p_K)$ of sample and a distribution $P_s = (p_1^s, p_2^s, \cdots, p_K^s)$ of a prior feature space of the speaker s , then filtering a possible solution set S_p containing a true solution: $\max_{l \in S_p} \{d_{sl}\} < \min_{l \notin S_p} \{d_{sl}\}$;

iv. calculating a distance measurement $D_s = \sum_t \sum_{k: g_k \in \Psi_t} (p_k - p_k^s)^\beta$ of the movement trajectory $\Psi_1 \Psi_2 \cdots \Psi_T$, selecting possible solutions $s^* = \arg \min_{s \in S_p} (D_s)$ from S_p , completing the speaker recognition.

Specifically, in the step 3), only the distribution information $P = (p_1, p_2, \cdots, p_K)$ of the space of the voice sample or the distance $D_s = \sum_t \sum_{k: g_k \in \Psi_t} (p_k - p_k^s)^\beta$ of the movement trajectory both can be used to obtain a good speaker recognition performance.

Compared with the prior art, the present invention has the following advantages and beneficial effects:

1. The present invention provides a method for speaker recognition based on trajectories in feature spaces of voice samples, wherein the establishment of the voice feature space is to cluster a large number of voice features, without the need for labelled data. The data samples for establishing the voice feature space can be

derived from different speakers. There are no exact requirements for the content of the speech, the age of the speaker, and the language. This overcomes the problem of requiring a large number of labelled voices in the neural network method, and that data collection established in the voice space is easy to realize.

2. The present invention provides a method for speaker recognition based on trajectories in feature spaces of voice samples, which is based on the location and trajectory information of the speaker voice feature in the voice feature space, and is different from the signal source generation model method, such as Hidden Markov Model (HMM) etc. The location is relative, while the generative model is absolute; compared with the deep neural network method, it is interpretable. Each piece of knowledge data has certain physical semantics. For example, the correlation degree distribution information $P = (p_1, p_2, \dots, p_K)$ of the sample feature in the space Ω expresses the active space range of the sample (the space represented by the identifier set corresponding to the non-zero element), and also expresses the distribution in this space.

3. The present invention provides a method for speaker recognition based on trajectories in feature spaces of voice samples, with concept that the voice features are located in space. For the voice features of different speakers, the positioning is performed on the established voice feature space, and the correlation degree is used to represent the voice feature positioning information of different speakers, which can be expressed with less calculation to show the distinction between different speakers, which has lower computational complexity than the GMM or HMM methods that require generative models to model each speaker.

4. The present invention provides a method for speaker recognition based on trajectories in feature spaces of voice samples, wherein the voice feature space identifier set is a reference system for locating the voice features of the speaker. It is a relative relationship and without strict relationship requirement with the sample to be recognized. Therefore, the feature space is shared. The established voice feature space can be transferred to other speaker data sets for recognition. For example, the voice feature space of a speaker of a language can be used as the voice feature space of a speaker recognition of another language, and data sharing is realized.

Description of the Figures

Figure 1 is a block flowchart of a speaker recognition method in Embodiment 1 of the present invention.

Figure 2 is a flowchart of steps for establishing a voice feature space in Embodiment 1 of the present invention.

Figure 3 is a flowchart of steps for generating distribution information and trajectory information of the voice feature space of a speaker in Embodiment 1 of the present invention.

Figure 4 is a flowchart of steps for recognizing a voice sample to be recognized in Embodiment 1 of the present invention.

Detailed Description

Hereinafter, the present invention will be further described in detail with reference to the examples and drawings, but the implementation of the present invention is not limited thereto.

Embodiment 1:

This embodiment provides a speaker recognition method based on the feature space trajectory of voice samples. A block flowchart is shown in Figure 1, and comprises the following three steps:

1) Establishing voice feature space Ω . Any clean voice sample can be used without any constraints on factors such as speaker, language, etc., and then the voice sample is clustered in the feature space using the clustering method. The sub-category data obtained by clustering is expressed as the expression of the voice feature space $\{g_k, k = 1, 2, \dots, K\}$;

2) Constructing speaker knowledge, comprising two parts of the distribution information in the voice feature space and movement trajectory information of the speaker;

3) For the voice samples to be recognized, use the distribution information of voice feature space and the movement trajectory information of the speaker of the voice sample for recognition.

Refer to Figure 2, which is a flowchart of steps for establishing a voice feature space in this embodiment, using voice data of speakers in the aishell Chinese corpus as the unlabelled voice sample set. There are a total of 400 speakers in aishell. 60 wav files of each person are selected to train the voice feature space, to extract 12-dimensional MFCC features of the unlabelled voice sample set $X = \{x_1, x_2, \dots, x_N\}$ to obtain the feature set $F_x = \{f_i^x, i = 1, 2, \dots, t_N\}$, where f_i^x is the short-term frame feature and t_N is the sum of the number of frames of all samples;

Then using $F_x = \{f_i^x, i = 1, 2, \dots, t_N\}$ feature sequence to train a GMM with a degree of mixing of K, discarding the weight information of the GMM, and retaining each Gaussian component as an identifier set of the voice feature space, wherein K is the number of audio feature space identifiers, and the number of identifiers K is selected to be 4096 in order to give a high-precision description of the audio feature space;

The voice feature space identifier is expressed as $\Omega = \{g_k, k = 1, 2, \dots, K\}$, where $g = N(m, U)$ is the multi-dimensional Gaussian distribution function;

Referring to Figure 3, it is a flowchart of steps for generating distribution information and trajectory information of the voice feature space of a speaker in this embodiment. For each person in aishell, 20 wav files are used to label the voice feature space. The target speaker voice sample set

$Y = \{(y_1, s_1), (y_2, s_2), \dots, (y_M, s_M)\}$, $s_i \in S = \{S_l, l = 1, 2, \dots, L\}$ (speaker set), the sample of speaker

S_l is $Y_l = \{y_m | s_m = S_l, m = 1, 2, \dots, M\}$, the audio feature sequence is extracted as

$F_l = \{f_1, f_2, \dots, f_{t_l}\}$; calculating the position correlation between all the features f_t of the voice sample and the space identifier $g_k(m_k, U_k)$:

$$p_t^k = \frac{1}{\left[(f_t - m_k)^T U_k^{-1} (f_t - m_k) \right]^\alpha}$$

Calculating the expected value of the position correlation between the speaker sample set and the space identifier $g_k(m_k, U_k)$:

$$B^k = \sum_n \sum_t p_{n,t}^k$$

where $p_{n,t}^k$ is the position correlation between the t -th frame feature of the n -th sample and the identifier $g_k(m_k, U_k)$;

Calculating the distribution of speaker feature space as:

$$p_k = \frac{B^k}{\sum_k B^k}$$

Processing the registered voice of each speaker in the target speaker set to obtain the distribution information of voice feature of each speaker.

The time sequence information of the movement trajectory of the voice sample is expressed as the neighborhood sequence $\Psi_1 \Psi_2 \cdots \Psi_T$ of the voice sample feature, and the δ neighborhood of the sample feature f_t is $\Psi_t = \{g_k | d_{tk} < \delta\}$, where d_{tk} is the Mahalanobis distance between the feature and the distribution, that is $d_{tk} = \sqrt{(f_t - m_k)^T U_k^{-1} (f_t - m_k)}$;

Referring to Figure 4, it is a flowchart of steps for recognizing voice samples in this embodiment. The feature of the voice sample to be recognized is $f = \{f_1, f_2, \cdots, f_T\}$, and the recognition process is as follows:

Calculating the distribution $P = (p_1, p_2, \cdots, p_K)$ of voice $f = \{f_1, f_2, \cdots, f_T\}$ in the feature space Ω ;

where the position correlation between the feature f_t and the space identifier $g_k(m_k, U_k)$ is:

$$p_t^k = \frac{1}{\left[(f_t - m_k)^T U_k^{-1} (f_t - m_k) \right]^\alpha}$$

The position correlation between the speaker sample to be recognized and the space identifier $g_k(m_k, U_k)$ is:

$$p_k = \frac{\sum_t p_t^k}{\sum_k \sum_t p_t^k}$$

Determining the trajectory $\Psi_1 \Psi_2 \cdots \Psi_T$ of the voice feature $f = \{f_1, f_2, \cdots, f_T\}$ in the feature space Ω , where $\Psi_t = \{g_k | d_{tk} < \delta\}$;

Calculating the distance $d_s = \|P - P_s\|^\beta = \sum_k (p_k - p_k^s)^\beta$ between the sample distribution

$P = (p_1, p_2, \cdots, p_K)$ and the distribution $P_s = (p_1^s, p_2^s, \cdots, p_K^s)$ of the prior feature space of speaker s, where α is 2, then filtering the possible solution set S_p containing the true solution:

$\max_{l \in S_p} \{d_{sl}\} < \min_{l \notin S_p} \{d_{sl}\}$, selecting the 10 speakers with the smallest distance as the candidate recognition results;

Calculating the distance measurement $D_s = \sum_t \sum_{k: g_k \in \Psi_t} (p_k - p_k^s)^\beta$ of the trajectory $\Psi_1 \Psi_2 \cdots \Psi_T$, where

α is 2, and choosing the speaker with the smallest trajectory distance from the 10 candidates as the recognition result, that is $s^* = \arg \min_{s \in S_p} (D_s)$.

Embodiment 2:

This embodiment provides a speaker recognition method based on trajectories in feature spaces of voice samples, comprising the following steps:

Step 1. using voice data of English corpus timit to establish a voice feature space identifier set;

Step 2. using voice data in the aishell corpus to register the target speaker set, as in Embodiment 1;

Step 3. recognizing voice samples to be recognized, as in Embodiment 1.

The recognition effect obtained is slightly different from that in Embodiment 1, which can prove that the voice feature space of a speaker of a language can be used as the voice feature space of a speaker recognition of another language, and data sharing is realized.

The above are only preferred embodiments of the present invention patent, but the scope of protection of the present invention patent is not limited to these. Equivalent substitutions or changes made by anyone familiar with this technical field according to the technical solution of the present invention patent and the concept of the present invention patent, within the scope disclosed by the present invention patent, all belong to the protection scope of the present invention patent.

Claims

1. A speaker recognition method based on trajectories in feature spaces of voice samples, wherein a voice sample can be regarded as a movement in a voice feature space, and has characteristics of an active space and a trajectory in a space, characterized in that, the method comprises the following steps:

step 1) constructing the voice feature space Ω : clustering unlabelled voice samples in the feature space using a clustering method, generating, based on subcategory data obtained from the clustering, a certain expression of the subcategory data as an expression of the voice feature space $\Omega = \{g_k, k = 1, 2, \dots, K\}$;

step 2) constructing speaker knowledge: using clean voice samples labelled with speaker attributes to obtain corresponding distribution information and movement trajectory information in the voice feature space Ω ;

step 3) speaker recognition: for a voice sample to be recognized, first obtaining a distribution expression and a trajectory in the voice feature space of the sample, then using the distribution information in the voice feature space of a speaker to calculate a difference between a sample distribution and a prior distribution, and a cumulative local distribution difference along the trajectory, as a basis for a speaker identification and to perform a determination.

2. A speaker recognition method based on trajectories in feature spaces of voice samples according to claim 1, characterized in that, in a process of constructing the voice feature space Ω of the step 1), any clean voice sample can be used, and there is no constraint on factors of speakers and languages.

3. A speaker recognition method based on trajectories in feature spaces of voice samples according to claim 1, characterized in that, the expression of the voice feature space $\Omega = \{g_k, k = 1, 2, \dots, K\}$ can be a distribution function of class data, a cluster center vector or a generative model, which are identifiers with positioning ability, namely, feature space identifiers, a class identifier size K used in the voice feature space determines a granularity of the expression of the voice feature space, the larger the K , the finer the expression of the voice feature space.

4. A speaker recognition method based on trajectories in feature spaces of voice samples according to claim 1, characterized in that, in the step 2), using the clean voice samples labelled with speaker attributes to label the voice feature space, when using a Gaussian distribution $g_k(m_k, U_k)$ as a space identifier, the distribution information in the feature space of the speaker is obtained as follows:

i. calculating a position correlation between each feature f_t of the voice sample and the space identifier $g_k(m_k, U_k)$, which is defined as:

$$p_t^k = \frac{1}{\left[(f_t - m_k)^T U_k^{-1} (f_t - m_k) \right]^\alpha}$$

in the formula, the space identifier is represented by a multi-dimensional Gaussian distribution, m_k represents a mean vector of the k-th Gaussian distribution, U_k represents a variance matrix of the k-th multi-dimensional Gaussian distribution;

ii. calculating an expected value of the position correlation between a speaker sample set and the space identifier $g_k(m_k, U_k)$:

$$B^k = \sum_n \sum_t p_{n,t}^k$$

in the formula, $p_{n,t}^k$ represents a correlation between the t-th frame feature of the n-th sample and the space identifier $g_k(m_k, U_k)$;

iii. calculating the distribution in the feature space of the speaker as:

$$p_k = \frac{B^k}{\sum_k B^k}.$$

5. A speaker recognition method based on trajectories in feature spaces of voice samples according to claim 4, characterized in that, in the step 2), a time sequence information of the movement trajectory of the speaker voice sample in the voice feature space Ω is expressed as a neighborhood sequence $\Psi_1 \Psi_2 \cdots \Psi_T$ of the voice sample feature, and a δ neighborhood of the voice sample feature f_t is $\Psi_t = \{g_k | d_{tk} < \delta\}$, where d_{tk} is the Mahalanobis distance between the voice sample feature and a distribution of the voice sample, that is $d_{tk} = \sqrt{(f_t - m_k)^T U_k^{-1} (f_t - m_k)}$.

6. A speaker recognition method based on trajectories in feature spaces of voice samples according to claim 5, characterized in that, a decision threshold of the δ neighborhood $\Psi_t = \{g_k | d_{tk} < \delta\}$ of the voice sample feature f_t refers to characteristics of a normal distribution, where $2 < \delta < 3$ is selected.

7. A speaker recognition method based on trajectories in feature spaces of voice samples according to claim 5, characterized in that, in the step 3), the speaker recognition process of the voice samples $f = \{f_1, f_2, \cdots, f_T\}$ comprises the following steps:

i. calculating the distribution $P = (p_1, p_2, \cdots, p_K)$ of voice sample $f = \{f_1, f_2, \cdots, f_T\}$ in the voice

feature space Ω , where $p_k = \frac{\sum_t p_t^k}{\sum_k \sum_t p_t^k}$;

ii. determining the movement trajectory $\Psi_1 \Psi_2 \cdots \Psi_T$ of the voice sample $f = \{f_1, f_2, \cdots, f_T\}$ in the voice feature space Ω , $\Psi_t = \{g_k | d_{tk} < \delta\}$;

iii. calculating a distance $d_s = \|P - P_s\|^\beta = \sum_k (p_k - p_k^s)^\beta$ between the distribution $P = (p_1, p_2, \cdots, p_K)$ of sample and a distribution of a prior feature space of the speaker s, then filtering a possible solution set S_p containing a true solution: $\max_{l \in S_p} \{d_{sl}\} < \min_{l \notin S_p} \{d_{sl}\}$;

iv. calculating a distance measurement $D_s = \sum_t \sum_{k: g_k \in \Psi_t} (p_k - p_k^s)^\beta$ of the movement trajectory

$\Psi_1 \Psi_2 \cdots \Psi_T$, selecting possible solutions $s^* = \arg \min_{s \in S_p} (D_s)$ from S_p , completing the speaker recognition.