(19) DANMARK

(10) DK/EP 2713367 T3

(12) Oversættelse af europæisk patentskrift

Patent- og Varemærkestyrelsen

(51) Int.Cl.: ***G 10 L 17/04 (2013.01)*** *G 10 L 17/10 (2013.01)* *G 10 L 17/16 (2013.01)* *G 10 L 17/24 (2013.01)*

(45) Oversættelsen bekendtgjort den: **2017-02-20**

(80) Dato for Den Europæiske Patentmyndigheds bekendtgørelse om meddelelse af patentet: **2016-11-09**

(86) Europæisk ansøgning nr.: **12006775.6**

(86) Europæisk indleveringsdag: **2012-09-28**

(87) Den europæiske ansøgnings publiceringsdag: **2014-04-02**

(84) Designerede stater: **AL AT BE BG CH CY CZ DE DK EE ES FI FR GB GR HR HU IE IS IT LI LT LU LV MC MK MT NL NO PL PT RO RS SE SI SK SM TR**

(73) Patenthaver: **Agnitio, S.L., Calle Gran Vía 39-8a planta, 28013 Madrid, Spanien**

(72) Opfinder: **Vaquero Avilés-Casco, Carlos, Calle de Constancia 27, 1° E, 28002 Madrid, Spanien**
Artiaga, Antonio Miguel, Dep. Ing. Electr. y Com., Edificio Ada Byron del campus Rio Ebro, EINA, Calle de María de Luna n°1, 50018 Zaragoza, Spanien
Buera Rodríguez, Luis, Calle del General Díaz Porlier 27, 5° A, 28001 Madrid, Spanien
García Gomar, Marta, Calle Nicaragua 10, 6° 5, 28936 Móstoles, Madrid, Spanien

(74) Fuldmægtig i Danmark: **Larsen & Birkeholm A/S Skandinavisk Patentbureau, Banegårdspladsen 1, 1570 København V, Danmark**

(54) Benævnelse: **Talegenkendelse**

(56) Fremdragne publikationer:
WO-A1-2007/131530
US-A1- 2006 111 905
US-A1- 2007 198 257
US-A1- 2010 131 273
MATSUI T ET AL: "Speaker adaptation of tied-mixture-based phoneme models for text-prompted speaker recognition", PROCEEDINGS OF THE INTERNATIONAL CONFERENCE ON ACOUSTICS, SPEECH, AND SIGNAL PROCESSING (ICASSP). SPEECH PROCESSING 1. ADELAIDE, APR. 19 - 22, 1994; [PROCEEDINGS OF THE INTERNATIONAL CONFERENCE ON ACOUSTICS, SPEECH, AND SIGNAL PROCESSING. (ICASSP)],, vol. i, 19 April 1994 (1994-04-19), pages I/125-I/128, XP010133577, DOI: 10.1109/ICASSP.1994.389339 ISBN: 978-0-7803-1775-8
VOGT R ET AL: "Experiments in Session Variability Modelling for Speaker Verification", ACOUSTICS, SPEECH AND SIGNAL PROCESSING, 2006. ICASSP 2006 PROCEEDINGS . 2006 IEEE INTERNATIONAL CONFERENCE ON TOULOUSE, FRANCE 14-19 MAY 2006, PISCATAWAY, NJ, USA,IEEE, PISCATAWAY, NJ, USA, 14 May 2006 (2006-05-14), page I, XP031330954, ISBN: 978-1-4244-0469-8
PATRICK KENNY ET AL: "Joint Factor Analysis Versus Eigenchannels in Speaker Recognition", IEEE TRANSACTIONS ON AUDIO, SPEECH AND LANGUAGE PROCESSING, IEEE SERVICE CENTER, NEW YORK, NY, USA, vol. 15, no. 4, 1 May 2007 (2007-05-01), pages 1435-1447, XP011177209, ISSN: 1558-7916, DOI: 10.1109/TASL.2006.881693

Fortsættes ...

DESCRIPTION

[0001] Speaker recognition systems can be used to confirm or refuse that a person who is speaking is who he or she has indicated to be (speaker verification) and can also be used to determine who of a plurality of known persons is speaking (speaker identification). Such a speaker identification system can be open-set as it is the possible that the speaker who is speaking is not one of the persons known to the system or close-set if the speaker is always in the set of the system. Such systems may find application in telephone banking, suspect identification and may generally be used in a security related context.

[0002] Such speaker recognition systems may require the user to say the same lexical content (e.g. the same key phrase) for both the enrolment and the recognition. Such a system is a text-dependent system, offering in some cases additional security because it requires recognising the identity of the speaker as well as the lexical content of the utterance.

[0003] Such recognition systems may also be text-independent, thus not setting any constraint with regard to the lexical content of the enrolment and of the recognition utterances. Such systems may have the advantage that people may be identified for example from common conversations, e.g. everyday conversations or enrolled with such common conversations of which files already exist.

[0004] Document US 2008/0312926 A1 discloses an automatic text-dependent, language-independent speaker voice-print creation and speaker recognition based on Hidden Markov Models (HMM) and automatic speech recognition (ASR) systems.

[0005] Document US 2007/0294083 A1 discloses a fast, text-dependent language-independent method for user authentication by voice based on Dynamic Time Warping (DTW).

[0006] Document US 6,094,632 discloses a speaker recognition device where the ASR system and speaker identification (SID) system outputs are combined.

[0007] Patrick Kenny provides an introduction to Speaker Verification related methods, in particular an algorithm, which may be used in Speaker Recognition systems in his article "Joint Factor Analysis of Speaker Session Variability: Theory and Algorithms".

[0008] Document US 2007/198257 A1 relates to speaker authentication.

[0009] Document W 2007/131530 A1 relates to a method for compensating intersession variability for automatic extraction of information from an input voice signal representing an utterance of a speaker.

[0010] Document US 2010/0131273 discloses a device, system, and method of liveness detection utilizing voice biometrics.

[0011] Another prior art document is the document "Support Vector Machines versus Fast Scoring in the Low-Dimensional Total Variability Space for Speaker Verification" by N. Dehak et al. in Interspeech, Brighton, London, Sept 2009.

[0012] It is known to use Hidden Markov Models (HMM) consisting of set of states which correspond to a deterministically observable event and are connected by transition probability arcs. States are defined on a vector of parameters and are extracted from the voice signal. Each state has an associated probability density function (pdf), which models the feature vectors associated to that state. Such a probability density function may for example be a mixture of Gaussian functions (Gaussian Mixtures Models, GMM), in the multi-dimensional space of the feature vectors, but other distributions may also be used.

[0013] The Hidden Markov Model is defined by the transition probabilities a_{ij} associated with the arcs representing the probability of moving from state i to state j, the initial state distributions p_i, which are associated to the states and are the initial probabilities of each state and the observation probability distribution b_i which is associated with the state i and may for example be a GMM. Those observation probability distributions are defined by a set of parameters depending on the nature of the distributions.

[0014] Conventional approximations for using Hidden Markov Models in text-dependent speaker recognition frameworks usually requires a transcription of the used phrase which is needed to build the speaker HMM by applying some kind of HMM adaption, for example, a maximum a posterior (MAP) (as disclosed in e.g. J. Gauvin and C. Lee "Maximum Posteriori Estimation for Multivariate Gaussian Mixture Observations of Markov Chains" IEEE Transactions on Speech and Audio Processing, 2(2): 291-298) or maximum likelihood linear regression (MLLR) (as disclosed e.g. in C.J Leggeter and P.C. Woodland in "Maximum likelihood linear regression for speaker adaptation of the parameters of continuous density Hidden Markov Models") or other

adaptations from a starting point model like a concatenation of generic HMMs representing units (e.g. phonems or words) of audio signals e.g. the phrase. In this framework, the generic HMMs are usually called Universal Background Model. From this, a scoring can be computed using a suitable algorithm like for example Viterbi or forward-backward algorithm as disclosed in L. R. Rabiner "a tutorial of Hidden Markof Models and selected applications in speech recognition", Proc. Of IEEE77 (2): 257-286, DOI:10.1109/5. 18626. [1].

[0015] Such generic HMMs usually require supervised training because every unit (e.g. phoneme, word, ...) needs to be associated with a certain HMM. From this the speaker recognition framework can be classified depending on how the transcription is obtained. Possibilities on how such a transcription can be obtained comprises prior knowledge, using conventional speech recognition systems or using universal speech recognition systems as described for example in US 2008/0312926. However, these approaches generally require supervised training and/or are computationally intensive, require a large amount of memory, are usually language dependent and/or are not very flexible. The classical approaches for text dependent HMM based speaker recognition systems may additionally have the disadvantage that the speaker HMM model has a direct relation with the transcription which may be stolen in at least one point of the system.

[0016] In classical speaker recognition using HMM adaption techniques, all the information of the feature vectors is incorporated into the speaker model, even though some information, like for example the channel, is not a typical feature of the speaker and should thus not be included in the speaker model.

[0017] For these reasons, classical text-dependent speaker recognition approaches have considerable limitations.

[0018] Some of their problems are the above described; storage of the transcription or an estimation of the transcription of the speaker phrase, the use of a speaker recognition or phonetic decoder making the system use a lot of memory and unsuitable for small devices like tablets or smart phones, and the fact that they do not compensate the channel or other negative effects of the speech signal.

[0019] Preferably, an improved system may take advantage of the information of the temporal sequence of the feature vectors and provide satisfactory performance and accuracy without using a transcription of the utterance, e.g. the speech phrase.

[0020] The present invention relates to automatic speaker recognition and solves at least one or more of the above mentioned problems. In particular, the invention relates to an automatic text-dependent, language-independent speaker recognition taking advantage of the temporal correlation of the speaker information of the feature vectors of the voice sample, but without incorporating the transcription or estimation thereof to any point of the system. Thus, satisfactory results may be achieved with low memory and computational time resources, so that it may be incorporated even in small devices, e.g. smart phones

[0021] The invention comprises a combination of classical text-independent approaches with a text-dependent approach in order to exploit the dynamic information. Effectively, classical text-independent approaches such as Joint Factor Analysis (JFA) as disclosed for example in the above mentioned article "Joint Factor Analysis of Speaker Session Variability: Theory and Algorithms" by Patrick Kenny or i-vector paradigms as explained for example in the document "Support Vector Machines versus Fast Scoring in the Low-Dimensional Total Variability Space for Speaker Verification" by N. Dehak et al. in Interspeech, Brighton, London, Sept 2009 do not use temporal correlation of voice samples, but very efficiently compensate channel effects to extract only speaker information. On the other hand, classical text-dependent approaches take advantage of temporal correlation thus taking into account a key point in text-dependent problems, but not compensating for channel effects satisfactorily. A combination of these paradigms can be carried at a system level, for example, by fusing a text-dependent speaker recognition system and a text-independent speaker recognition system or by incorporating the text-independent strengths to a text-dependent framework.

[0022] The invention comprises a method according to independent claim 1. Favourable embodiments are disclosed in the dependent claims 2-13.

[0023] In particular, the invention comprises a method for text-dependent speaker recognition using a speaker model obtained by adaptation of a Universal Background Model wherein the speaker model is a speaker adapted Hidden Markov Model comprising channel correction. Such a method may comprise two parts, namely enrolment, where the model is adapted and the recognition part in which a speaker is verified or identified and an audio test is compared against a set of models or one model for verification.

[0024] A generic Universal Background Model (generic UBM) is usually adapted to a speaker after a speaker enrols.

[0025] It may be adapted based on one or more utterances (enrolment utterances) of the speaker. From such an utterance (e.g.

a phrase, a word, a fragment which is usually present as audio file or information) feature vectors (enrolment feature vectors) may be extracted. The utterances comprise speaker information and channel information, wherein the channel is everything between the speaker and the recording support, e.g. comprising a microphone through which the utterances have been transmitted, and anything else through which the utterances have been transmitted, e.g. cables, loudspeakers, etc.

[0026] This adaptation may be done in an unsupervised manner. In particular, no lexical content of the utterances or approximation thereof may be necessary. Thus, no speech recognition or other transcription method may be necessary.

[0027] A text-dependent model for the corresponding utterance(s) may then be built by adapting the generic Universal Background Model of the text-dependent system with all the enrolment feature vectors of the corresponding utterance(s) (wherein usually each utterance comprises the same lexical content) and speaker. The enrolment feature vectors are usually extracted from the utterances provided for the enrolment. Channel correction may be applied such that the speaker model depends only on the speaker but not on the channel (as e.g. shown in the document by Kenny, mentioned above).

[0028] For each speaker, several models may be generated. Usually one speaker model is adapted from the UBM for one lexical content, e.g. a phrase, a word or a fragment. Each model may be obtained with one, two or more utterances, e.g. five utterances of the same lexical content.

[0029] In particular, once some utterances of one speaker are present as audio file or other audio information, e.g. of a certain phrase, a word, a fragment or something similar, short time feature vectors may be extracted. In such time feature vectors, for example Mel Frequency Cepstral Coefficients (MFCC) as shown e.g. by Davis, S. B. and Mermelstein, P. in "Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences", IEEE Trans. on Acoustic, Speech and Signal Processing, 28(4): 357-366, 1980 may be used. In the time feature vectors all the relevant information of the speaker, lexical content and other aspects (also undesired, e.g. channel effects) are compressed in a small vector at every time interval. For example, at every 10 milliseconds a 60 dimensional feature vector may be created.

[0030] With this information, a Joint Factor Analysis paradigm model as disclosed in the above mentioned article by Kenny may be applied to adapt a generic UBM, which is a HMM, compensating at the same time undesirable aspects of the voice signal like for example the channel. For a given speaker and channel h the complete model is:

$$M_h(s)=m + v\,y(s)+u\,x_h(s)+d\,z(s)$$

wherein m is a supervector obtained by concatenation of all the mean vectors of the generic UBM, $M_h(s)$ is the supervector modelling the speaker s and channel h, v is the rectangular eigenvoice matrix, y(s) is the hidden variable which includes the speaker information (speaker factors), u is the rectangular eigenchannel matrix, $x_h(s)$ is the hidden variable which includes the channel information (channel factors), d is a diagonal matrix for modelling those aspects of the voice signal which are not included in u and v, and z(s) is a hidden residual variable associated with d. The term d z(s) may be modelled as zero (and thus not taken into account).

[0031] For this adaption zero and first order statistics for the utterances which are used for the enrolment for the speaker adaptation of the UMB (enrolment utterances) are required. Some suitable algorithm like for example a forward-backward algorithm or a Viterbi algorithm may be carried out for the generic UBM and each of the utterances of the corresponding phrase and speaker.

[0032] For the enrolment utterances, the speaker usually has to say the same lexical content (e.g. a phrase, a word or a fragment) more than once, in particular more than 2, in particular more than 3 or in particular more than 5 times. For verification (recognition) in such systems (explained below) one utterance may be sufficient.

[0033] For each speaker, several models may be created (in particular adapted from a generic UBM) with different lexical contents. For each utterance with a lexical content for recognition of the speaker, usually the model corresponding to said lexical content is used.

[0034] For example, using a forward-backward algorithm, the probability of being in the Gaussian k of state i at time t $\gamma_{i,k}(t)$ can be calculated. The Annex shows how these probabilities can be calculated.

[0035] From this, the zero and first order statistics over time are:

$$N_{i,k}=\Sigma_t \gamma_{i,k}(t)$$

$$F_{i,k}=\Sigma_t \gamma_{i,k}(t)\, x_t$$

[0036] Herein x_t is the enrolment feature vector at time interval t.

[0037] From this, speaker and channel factors may be extracted and the mean vectors of the adapted HMM may be given as
$\overline{M_{lc}}(s) = m + v\ y(s)$

[0038] This equation may represent the means of the speaker adapted model (speaker model) and may also be used in a recognition process later on. The index lc refers to the lexical content, for which $M_{lc}(s)$ is adapted. y(s) may be calculated from the enrolment utterance(s) as shown e.g. in the document by Kenny, while v (like u and d, if required) may be obtained in a previous developing phase (after the unsupervised training of the generic UBM) as shown e.g. in the document by Kenny.

[0039] For practical purposes, the system may not store $M_{lc}(s)$, but may only store y(s). This may be advantageous as the vector y(s) may have considerably fewer dimensions than $M_{lc}(s)$. Because v and m may have to be stored anyway, the storage of y(s) instead of $M_{lc}(s)$ may reduce the necessary system resources.

[0040] The transition probabilities may be adapted from the generic UBM in order to complete the speaker model (speaker adapted model), which may have the same parameters as the UBM except the means, which may be computed using the equations provided before and the transition probabilities as provided below. The transmission probabilities of the generic HMM may be adapted from the known transition probabilities of the generic HMM and may be given as

$$\hat{a}_{ij} = \frac{\sum_t \tau_t(i,j)}{\sum_j \sum_t \tau_t(i,j)}$$

$$\tau_t(i,j) = \frac{\alpha_t(t) a_{ij}\ b_j(x_{t+1})\ \beta_j(t+1)}{\sum_j \alpha_t(i) a_{ij}\ b_j(x_{t+1})\ \beta_j(t+1)}$$

[0041] The meaning of the variables in these equations may be as explained e.g. in the annex.

[0042] In some embodiments of a method according to the invention, only mean vectors and transition probabilities of the generic UBM may be adapted in the speaker model as they have proven to be the most important parameters. In other embodiments, all of the parameters of the generic UBM may be adapted to the speaker. At this point, enrolment may be completed.

[0043] In a method according to the invention, the generic Universal Background Model may be trained in an unsupervised training before it is adapted for a certain speaker and text content. In particular, it may be trained from a set of audio information without the information of the transcriptions using a suitable algorithm like for example Expectation Maximization algorithm (EM) (as disclosed e.g. in A. P. Dempster, N. M. Laird and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm", Journal of the Royal Statistical Society, 39(1)) or Variational Bayes algorithm (VB) (as disclosed e.g. in C. M. Bishop, "Pattern Recognition and Machine Learning", Springer Verlag).

[0044] For this unsupervised training, utterances of several speakers may be used, for example, of more than five, in particular of more than ten, in particular of more than twenty speakers which may speak more than one, in particular more than five, in particular more than ten, and in particular more than fifteen different languages and/or utterances of more than 1, in particular more than 2, in particular more than 5, in particular more than 10 and in particular more than 20 Hungarian speakers may be used for the unsupervised training of the generic Universal Background Model. (Hungarian is considered to cover almost all of the possible sounds that can be pronounced in most of the main spoken languages.) Of one, more or all speakers whose utterances are used more than one, in particular, more than 5 utterances may be used.

[0045] In such a training phase, preferably as many as possible free speech conversations are used to train the generic Universal Background Model to be ready for any utterance, e.g. password, phrase or language that may be used during operation.

[0046] Such an unsupervised training may be advantageous because that way each state of the generic Universal Background Model does not have a physical meaning (e.g. is not confined to a certain phoneme) thus increasing the security and reducing the cost necessary to create such a generic Universal Background Model.

[0047] The topology of the generic Universal Background Model may be selected to comprise a transition probability from each possible state to itself and each possible other state thus making sure that each possible utterance, e.g. password, phrase or language can be used in that generic Universal Background Model.

[0048] In addition, the observation probability distributions may be Gaussian mixture models optionally comprising diagonal covariance matrices and four components.

[0049] In a method according to the invention, the number of states may be fixed.

[0050] In particular, the number of states may be set to less than thirty states, to thirty or to more than thirty, or the number states may set to the number of Hungarian phonemes, or a number which may be estimated by an analysis of the spectral properties of a speech signal.

[0051] This may, e.g. be done when a precise generic UBM for a certain lexical content, e.g. a certain phrase, a certain word or a certain fragment is needed. Such a generic UBM may have to be retrained for every new lexical content, e.g. phrase, word or fragment.

[0052] In particular, it may be assumed that as many states may be needed as spectral changes are detected in a signal. An algorithm to find the number of spectral changes may be based on a short time smooth linear prediction analysis. Given two adjacent segments, f and g, they may be analyzed in order to determine whether there is a spectral change between them or not. A possible divergence measure can be computed by

$$D(f,g) = \log\left(\frac{a_g^{tt} R_{ff}\, a_g}{a_f^{tt} R_{ff}\, a_f}\right)$$

[0053] If the divergence measure is higher than a certain threshold, it may be assumed that the spectral change was present between segments f and g. R_{ff} may be the autocorrelation signal of the segment f, and a* (wherein * may be f or g) may be the filter coefficients of the optimum linear prediction filter for * for example extracted from the autocorrelation signal by solving the minimum squared prediction error problem. This may be similar to the voice activity detector (VAD) of GSM (ETSI EN 300 730).

[0054] The method may further comprise adapting one or more parameters, e.g. a number of states and/or e.g. the generic Universal Background Model to a lexical content, e.g. a passphrase or password. Thus, if a certain lexical content is used and already known, it may not be necessary to cover all sounds with the generic Univeral Background Model. In particular, the rest of the generic Universal Background Model parameters may then be found using data and classical HMM adaption algorithms such as e.g. MAP. For this, the data of about 10, e.g. 5-15 speakers speaking the lexical content (e.g. a (pass)phrase, (pass)word or a fragment) a few times, e.g. 2-5 times, may be used and may already be sufficient.

[0055] Following that, eigenchannel matrix u and eigenvoice matrix v may be trained in a development session before the generic UBM may be adapted for a speaker. By having these two matrices channel correction may be performed since the channel information (eigenchannel matrix) is kept separate from the speaker information (eigenvoice matrix), which may be processed separately. In this development session in the matrix d, which may model the aspects of the voice signal not included in u and v, may also be trained. In other embodiments, d may be modeled as 0, thus not being present and not needing to be trained.

[0056] A method according to the invention may also comprise the step of verifying, whether a test signal was spoken by the target person. Such a verifying may be done in an unsupervised manner. No transcription or approximation of the lexical content may therefore be necessary for the verifying step, and in particular, e.g. no speech recognition or transcription method may be necessary.

[0057] For this, a testing feature vectors $\{x_1, x_2, x_3 \ldots x_T\}$ may be created from the spoken test signal.

[0058] In particular, for example MFCC may be used wherein in every 10 milliseconds a 60 dimensional feature vector is created.

[0059] Furthermore, the speaker model represented by the speaker factors y(s) and zero and first order statistics computed with the testing feature vectors may be used for the calculation.

[0060] Statistics may be computed using the generic UBM and the speaker model with a suitable algorithm, e.g. forward-backward or Viterbi algorithm. Then the log likelihoods of the speaker model LL_{spk} and the generic UBM model LL_{UBM} may be computed.

[0061] In particular, a suitable algorithm, for example, the Viterbi or forward-backward algorithm may be carried out with the testing feature vectors and the speaker adapted HMM, wherein the means and the transition probabilities may be the adapted ones.

[0062] Such a suitable algorithm like for example the Viterbi or forward-backward algorithm may then provide a most likely path for the feature vectors (testing feature vectors) over the speaker adapted HMM states in this case $q=\{q_1,q_2,...q_T\}$.

[0063] Then the probability of being in Gaussian k of state i of the generic UBM at time t may calculated as

$$\gamma_{i,k}^{Vit}(t)=\begin{cases}\dfrac{w_{i,k}\ b_{i,k}(x_t;\mu_{i,k},\sum_{i,k})}{\sum_k w_{i,k}\ b_{i,k}(x_t;\mu_{i,k},\sum_{i,k})} & if\ i=q_t\\ 0 & otherwise\end{cases}$$

and the zero and first order statistics may be given as

$$N_{i,k}=\sum_t \gamma_{i,k}^{Vit}(t)$$

$$F_{i,k}=\sum_t \gamma_{i,k}^{Vit}(t)x_t$$

[0064] Herein x_t is the feature vector at time interval t and $\mu_{i,k}$ and $\sum_{i,k}$ are the mean and covariance matrix of Gaussian k of state i of the generic UBM.

[0065] Then a log likelihood of the feature vectors regarding the speaker adapted HMM may be obtained e.g. disclosed in the document by Kenny (LL_{SPK}). Here, the speaker adapted model may be used only to determine the (most likely) path, but not to compute the statistics.

[0066] This may be in contrast to the classic JFA approach using the generic model to determine a (most likely) path, producing different misalignment, especially in text-dependent approaches. However, the classic JFA approach may also be taken.

[0067] Then a corresponding step may be carried out with a generic UBM and the testing feature vectors, thus leading to a most likely path

$Q^{UBM}=\{q_{UBM1}, q_{UBM2}...q_{UBMT}\}$.

For this again, a suitable algorithm like for example the Viterbi or forward-backward algorithm may be used.

[0068] The probability of being in Gaussian state k of the generic UBM can be computed as

$$\gamma_{i,k}^{UBM,Vit}(t)=\begin{cases}\dfrac{w_{i,k}\ b_{i,k}(x_t;\mu_{i,k},\sum_{i,k})}{\sum_k w_{i,k}\ b_{i,k}(x_t;\mu_{i,k},\sum_{i,k})} & if\ i=q_{tUBM,t}\\ 0 & otherwise\end{cases}$$

with corresponding zero in first order statistics:

$$N_{i,k}^{UBM}=\sum_t \gamma_{i,k}^{UBM,Vit}(t)$$

$$F_{i,k}^{UBM}=\sum_t \gamma_{i,k}^{UBM,Vit}(t)x_t$$

[0069] Again x_t is a feature vector in the time interval t. Then the log likelihood ratio of the feature vectors may be calculated (e.g. as disclosed in the document by Kenny (LL_{UBM})).

[0070] As a final measure for the comparison between the testing signal (utterance), for example, an audio file with an utterance, e.g. testing phrase or a password spoken by a speaker with regard to one selected speaker model has the log likelihood ratio (LLR_{td}) computed as

$$LLR_{td}=LL_{spk}-LL_{UBM}+\sum_{i=1}^{T-1}\log(\hat{a}_{q_i q_{i-1}})-\sum_{i=1}^{T-1}\log(a_{q_{UBM,i}q_{UBM,i+1}})$$

[0071] In this case, the log likelihood ratio may in addition to the logs of the likelihood (LL_{spk} and LL_{UBM}) also comprise the logarithm of terms describing the speaker adapted transition probabilities and generic transition probabilities over the most likely paths of the corresponding models which have been calculated. This may be advantageous as it may take the temporal progression along the HMMs into account.

[0072] Such a method may further comprise a step of identifying a target person by identifying the speaker of a speaker model adapted from the generic Universal Background Model with the highest likelihood score before verifying whether the target person is indeed the one to have spoken the test signal as explained before.

[0073] An example of another implementation may also comprise a method for text-dependent speaker recognition using a combination of a text-dependent and a text-independent system, wherein the text-dependent system is adapted and wherein in addition a model of the text-independent system is also built for the speaker and the phrase.

[0074] The adaptation may optionally be done in an unsupervised manner (unsupervised way). In particular, no lexical content of the utterances or approximation thereof may be necessary. Thus, no speech recognition or other transcription method may be necessary.

[0075] The text-dependent model may use a generic UBM, e.g. a HMM as starting point, In particular, it may be a UBM which was trained in an unsupervised way and may have a topology as explained above. Thus, no transcription of the lexical content may be needed for the training and no transcription of the lexical content may be needed for the following adapting to a speaker. The generic UBM may for example be a HMM and may be adapted to the speaker for example by a suitable algorithm like MAP or a Bayesian algorithm (as disclosed e.g. in C.M Bishop "Pattern Recognition and Machine Learning", Springer Verlag).

[0076] When a text-dependent and text-independent speaker recognition systems are combined, the text-independent one may be composed by classic JFA or i-vector framework, and it may give a score or a log likelihood ratio with channel compensation. On the other hand, the text-dependent system may not do any kind of channel compensation, and it may give a score or a log likelihood ratio.

[0077] The procedure may be as following:

For the procedure, e.g. feature vectors as mentioned above may be used, e.g. with the parameters of 10 ms time intervals and 60 dimensional MFFC.

[0078] For text-dependent system, the generic UBM, which may be a HMM, may be adapted to the speaker with some enrolment audio by MAP, using a suitable algorithm, like for example, using Viterbi or forward backward algorithm. Then, the enrolment may be finished. For recognition, a suitable algorithm, like e.g. Viterbi may be applied with the testing audio and the speaker model, getting LLR_{spk}. Also, the suitable algorithm, like e.g. Viterbi may be applied with the testing audio and the generic UBM, getting LLR_{UBM}. No channel compensation may be necessary and thus channel compensation may not be done. Finally, the LLR_{td} may be computed as
$LLR_{td}=LLR_{spk}-LLR_{UBM}$.

[0079] For the text-independent system, the generic UBM, which may be a GMM, may be used to generate the speaker model using JFA or i-vector framework. Then, the enrolment may be finished. Here, channel compensation may be done. For recognition, JFA or i-vector framework may be applied, thus allosing computing of LLR_{ti}.

[0080] Finally, the final log likelihood ratio (LLR) may be obtained as:
$LLR=\alpha\ LLR_{td} + \beta\ LLR_{ti} + \delta$.

[0081] The enrolment audio and testing audio may be the same for both systems, so the user may have to say the same lexical content, e.g. (pass)phrase for enrolment and recognition.

[0082] In such a method, the transition probabilities and the means of the probability density distribution which may for example be GMM, of the generic UBM of the text-dependent system, which may be a HMM, may be modified in such an approach. Those

parameters have been proven to be the most important ones. If an MAP adaption is used, labelling of each adaption feature vector associating one state is required. Such labelling may be done by a suitable algorithm, like the forward-backward algorithm or Viterbi algorithm.

[0083] For building the model of the text-independent system for the corresponding phrase and speaker, a suitable approach may also be chosen, such as for example, joint factor analysis or i-vector paradigms as disclosed by Dehak et al. In such a building, the same feature vectors may be used as for adapting the model of the text-dependent system.

[0084] As a generic UBM for the text-dependent system, GMM may be used, however the use of HMM is also possible in other embodiments.

[0085] Such a method may also comprise the step of verifying whether a test signal was spoken by the target person. In particular, such a verifying step may comprise extracting short time feature vectors from the test signal, e.g. for example from a test phrase or password. Those feature vectors may for example be extracted by MFCC where the parameters may e.g. be a time distance of 10 milliseconds with a 60 dimensional feature vector. Given those feature vectors, a log likelihood ratio of the speaker adapted and non-adapted (generic UBM) model may be computed for the text-independent system (LLR_{ti}). Such a verifying may be done in an unsupervised manner which may have the advantages described before with regard to an unsupervised manner in a verifying step.

[0086] Following that, the feature vector of the one speaker may also be compared against the generic model of the text-dependent system (with which LL_{UBM} is found), for example; a HMM and the speaker model of the text-dependent system (with which LL_{SPK} is found) previously built using a suitable algorithm like for example MAP or MLLR using e.g the forward-backward algorithm or Viterbi algorithm. From the two log likelihoods, a combined log likelihood ratio for the text-dependent models may be calculated as:

$LLR_{td} = LL_{spk} - LL_{UBM}$

[0087] The final log likelihood ratio (LLR) may be obtained as a linear combination of the text-dependent and the text-independent log-likelihood ratios to which an independent term is also added. It may be given as:

$LLR = \alpha\, LLR_{td} + \beta\, LLR_{ti} + \delta$

wherein LLR_{td} is the log likelihood ratio of the text-dependent system and LLR_{ti} is the log likelihood ratio of the text-independent system and δ is an independent term. The scalar values α and β which are the coefficients for the text-dependent log likelihood ratio and the text-independent log likelihood ratio and independent term δ may be trained during a development session with external data, which is usually independent of the speaker. The term δ may be used because of the physical meaning of LLR.

[0088] The LLR may be used in forensic scenarios, for example, LLR may represent the ratio between the hypothesis that the two audios one is comparing have been spoken by the same person and the hypothesis that both audios have not been spoken by the same person.

[0089] Then, LLR higher than 0 may mean that it is more probable that both audios have been spoken by the same person that they have not been spoken by the same person.

[0090] However, it may not be simple and the threshold may not usually be 0 because one may have to consider the priori information which may take into account some other information that may not be the voice, and the penalties of having an error (false acceptance and false rejection). Because of that it may be important that the LLR may be well-calibrated. In that case a threshold may be fixed considering the priori information and the costs of the errors very easily.

[0091] Because of that, δ may be required and trained. In some other embodiments, δ may be set to 0. Such a linear combination of the two log likelihood ratios of the text-dependent and the text-independent system allow with the appropriate training to achieve better results than when using a text-dependent or a text-independent system only. Again, for the verification, the final log likelihood ratio LLR may be compared against a threshold.

[0092] Such a method may also comprise the step of identifying a target person by identifying the speaker adapted model with the highest likelihood score as possible target speaker before verifying that the test signal was spoken by that speaker.

[0093] The invention also comprises a computer readable medium according to claim 14.

[0094] The invention further comprises a system comprising means according to claim 15.

[0095] The invention is further explained with the figures, wherein

Fig. 1
shows state of the art text-dependent systems; and
Fig. 2
shows a system comprising aspects of the invention.

[0096] In Fig. 1, a prior art speaker recognition system is explained.

[0097] Starting from an audio signal a transcription of the text is extracted. Following that, a starting point HMM may be created from the generic UBM using the transcription or an estimation of the transcription. Then, a speaker adapted HMM may be generated using the audio signal and the starting point HMM. For the adaptation a suitable algorithm like MAP or MLLR may be used which may further use one or more suitable algorithms thus, e.g., in some cases Viterbi or forward-backward algorithm may be needed also.

[0098] Following that, during the testing phase, a suitable algorithm like for example Viterbi and/or forward-backward algorithms may be used for deciding whether the target person is indeed the one who has spoken the test signal. Such a decision may be done by comparing a score against a set threshold.

[0099] Fig. 2 shows one system according to one embodiment of the invention. In particular, starting out from a generic UBM, the generic UBM, which may be a HMM, may be adapted using one or more utterances, e.g. audio signals. Thus a speaker adapted HMM may be generated. As can be seen, in such a system, no transcription of the content of the audio file may be necessary.

[0100] After the enrolment phase, testing may be done using the speaker adapted HMM in combination with a testing audio file and also using the generic UBM. A decision may then be made about the verification comparing the log likelihood ratio of the modified speaker model against the non-modified generic UBM against a threshold.

[0101] Figure 2 may represent a text-dependent solution wherein the channel is compensated or a text-dependent solution wherein the channel is not compensated. However, Figure 2 does not show a text-independent solution which may also be used in some embodiments of the invention.

[0102] In particular, in some embodiments, wherein the channel is not compensated in the text-dependent system, a fusion of the score with the score of the text-independent system may be necessary (not shown).

Annex:

[0103] Given the feature vectors for one utterance $X=\{x_1, x_2, x_3 ...\}$ and one HMM defined by

- $A = \{a_{ij}\}$, the transition probabilities, which are associated to the arcs and represent the probability of "moving" from state *i* to state *j*.
- $\Pi = \{p_i\}$, the initial state distribution, which are associated to the states and are the initial probabilities of each state.
- b_i, Observation probability distribution, which is associated to the state i and is a GMM, defined by:
  - $w_{i,k}$: a priori probability of Gaussian k of state i.
  - $b_{i,k}(x) = N\left(x; \mu_{i,k}, \sum_{i,k}\right)$:
    where $\mu_{i,k}$ and $\sum_{i,k}$ are the mean and covariance matrix of Gaussian k of state i.

[0104] We can define the probability of producing the sequence $\{x_1, x_2, x_3 ..., x_{t-1}\}$ while ending up in state i at time t as $\alpha_i(t)$:

$$\alpha_i(t)=b_i(x_t)\sum_j a_{ji}\alpha_j(t-1)$$

where

$$b_i(x_t)=\sum_k w_{i,k}b_{i,k}(x_t)$$

is the likelihood of the state i for x_i. We can define the probability of producing the sequence $\{x_t, x_{t-1}, ..., x_T\}$, where T is the total number of feature vectors, given that we are at state i at time t as $\beta_i(t)$:

$$\beta_i(t)=\sum_j a_{ij}\beta_j(t+1)b_j(x_{t+1})$$

[0105] Finally, the probability of being in Gaussian k of state i at time t is $\gamma_{i,k}(t)$:

$$\gamma_{i,k}(t)=\frac{\beta_i(t)w_{i,k}b_{i,k}(x_i)\sum_j a_{ji}\alpha_j(t-1)}{\sum_j \alpha_j(t)\beta_j(t)}$$

[0106] The initialization values for use in an algorithm, e.g. forward-backward algorithm, may be $\alpha_i(1)=p_i$ and $\beta_i(T+1)=1$.

REFERENCES CITED IN THE DESCRIPTION

This list of references cited by the applicant is for the reader's convenience only. It does not form part of the European patent document. Even though great care has been taken in compiling the references, errors or omissions cannot be excluded and the EPO disclaims all liability in this regard.

Patent documents cited in the description

- US20080312926A1 [0004]
- US20070294083A1 [0005]
- US6094632A [0006]
- US2007198257A1 [0008]
- USW2007131530A1 [0009]
- US20100131273A [0010]
- US20080312926A [0015]

Non-patent literature cited in the description

- **N. DEHAK et al.**Support Vector Machines versus Fast Scoring in the Low-Dimensional Total Variability Space for Speaker VerificationInterspeech, Brighton, London, 2009, [0011] [0021]
- **J. GAUVINC. LEE**Maximum Posteriori Estimation for Multivariate Gaussian Mixture Observations of Markov ChainsIEEE Transactions on Speech and Audio Processing, vol. 2, 2291-298 [0014]
- **L. R. RABINER**a tutorial of Hidden Markof Models and selected applications in speech recognitionProc. Of IEEE, vol. 77, 2257-286 [0014]
- **PATRICK KENNY**Joint Factor Analysis of Speaker Session Variability: Theory and Algorithms, [0021]
- **DAVIS, S. B.MERMELSTEIN, P.**Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken SentencesIEEE Trans. on Acoustic, Speech and Signal Processing, 1980, vol. 28, 4357-366 [0029]
- **A. P. DEMPSTERN. M. LAIRDD. B. RUBIN**Maximum likelihood from incomplete data via the EM algorithmJournal of the Royal Statistical Society, vol. 39, 1 [0043]
- **C. M. BISHOP**Pattern Recognition and Machine LearningSpringer Verlag [0043]

PATENTKRAV

1. Fremgangsmåde til tekstafhængig talegenkendelse under anvendelse af en talemodel opnået ved adaptation af en universel baggrundsmodel; hvor talemodellen er en taleadapteret skjult Markov-model, der omfatter kanalkorrektion, hvor fremgangsmåden omfatter at adaptere en generisk universel baggrundsmodel til et leksikalsk indhold og at træne en egenkanalmatrix og en egenstemmematrix i en udviklingssession, inden den generiske universelle baggrundsmodel adapteres til en talende.

2. Fremgangsmåde til talegenkendelse ifølge krav 1, hvor den universelle baggrundsmodel uovervåget adapteres baseret på indtegnede ytringer fra den talende, hvor statistik ekstraheres ved at anvende indtegnede trækvektorer, den universelle baggrundsmodel og en fremadgående-tilbagegående eller Viterbi algoritme.

3. Fremgangsmåde til talegenkendelse ifølge krav 1 eller 2, hvor kun middelvektorer og overgangssandsynligheder adapteres i talemodellen, eller hvor alle parametre adapteres i talemodellen.

4. Fremgangsmåde til talegenkendelse ifølge et af kravene 1 til 3, hvor den universelle baggrundsmodel i det tekstafhængige system trænes i uovervåget træning, inden den adapteres.

5. Fremgangsmåde til talegenkendelse ifølge et af kravene 1 til 4, hvor ytringer fra mere end 10 talende, som taler mere end 5 forskellige sprog eller flere end 5 ungarske talende anvendes til en uovervåget træning af den universelle baggrundsmodel i det tekstafhængige system.

6. Fremgangsmåde til talegenkendelse ifølge et af kravene 1 til 5, hvor topologien af den universelle baggrundsmodel i det tekstafhængige system vælges til at

omfatte en overgangsmulighed fra hver mulig tilstand til sig selv og enhver mulig anden tilstand.

7. Fremgangsmåde til talegenkendelse ifølge et af kravene 1 til 6, hvor observationssandsynlighedsfordelingerne er Gaussiske blandingsmodeller med diagonale kovariansmatrixer og fire komponenter.

8. Fremgangsmåde til talegenkendelse ifølge et af kravene 1 til 7, hvor antallet af tilstande er fastsat.

9. Fremgangsmåde til talegenkendelse ifølge et af kravene 1 til 8, hvor antallet af tilstande sættes til mindre end 30, 30, flere end 30, antallet af ungarske fonemer eller et antal skønnet ved en analyse af spektralegenskaberne ved et signal.

10. Fremgangsmåde til talegenkendelse ifølge et af kravene 1 til 9, der endvidere omfatter trinet at verificere på en uovervåget måde, om et prøvesignal blev talt af en målperson.

11. Fremgangsmåde til talegenkendelse ifølge et af kravene 1 til 10, hvor den taleadapterede model kun anvendes til at bestemme den mest sandsynlige sti, men ikke til at beregne statistikken.

12. Fremgangsmåde til talegenkendelse ifølge et af kravene 1 til 11, hvor det at verificere, om prøvesignalet blev talt af målpersonen, omfatter at beregne forskellen mellem de to begreber af den logaritmiske sandsynlighed af prøveaudioen og talemodellen og det logaritmiske produkt af overgangssandsynlighederne af den mest sandsynlige sti opnået med talemodellen og den logaritmiske sandsynlighed af prøveaudioen og den generiske universelle baggrundsmodel og det logaritmiske produkt af overgangssandsynlighederne af den mest sandsynlige sti opnået med den generiske universelle baggrundsmodel.

13. Fremgangsmåde til talegenkendelse ifølge et af kravene 1 til 12, hvor fremgangsmåden endvidere omfatter et trin med at identificere en målperson ved at identificere den taleadapterede model med den højeste sandsynlighedsscore.

14. Computerlæsbart medium, der omfatter computerlæsbare instruktioner til at udføre en fremgangsmåde ifølge et af kravene 1 til 13.

15. System, der omfatter midler til at udføre en fremgangsmåde ifølge et af kravene 1 til 13.

DRAWINGS

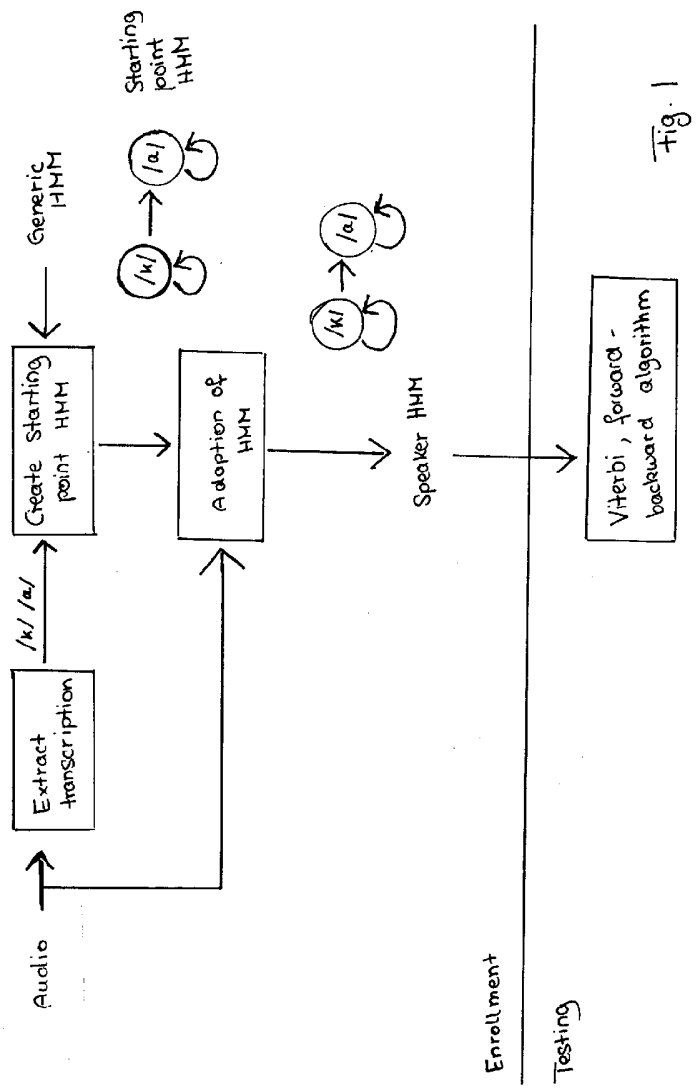


Fig. 1

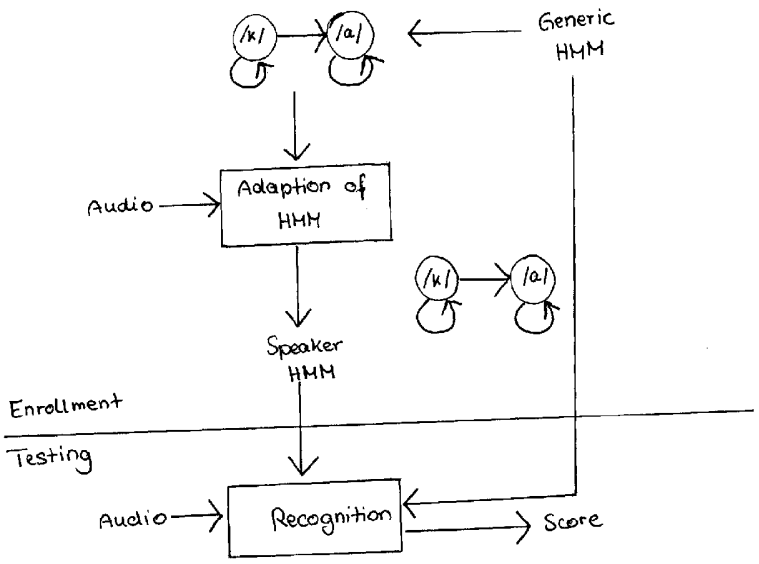

/k/
/a/
Generic HMM
Audio
Adaption of HMM
/k/
/a/
Speaker HMM
Enrollment
Testing
Audio
Recognition
Score

Fig. 2