BESCHWERDEKAMMERN DES EUROPÄISCHEN PATENTAMTS	**BOARDS OF APPEAL OF THE EUROPEAN PATENT OFFICE**	**CHAMBRES DE RECOURS DE L'OFFICE EUROPEEN DES BREVETS**

Internal distribution code:
(A) [ ] Publication in OJ
(B) [ ] To Chairmen and Members
(C) [ ] To Chairmen
(D) [X] No distribution

Datasheet for the decision of 4 June 2008

Case Number:	T 1546/05 - 3.4.03
Application Number:	96118549.3
Publication Number:	0779602
IPC:	G07C 9/00
Language of the proceedings:	EN

Title of invention:
Method and apparatus employing audio and video data from an individual for authentification purposes

Applicant:
AT & T Corp.

Opponent:
-

Headword:
-

Relevant legal provisions (EPC 1973):
EPC Art. 56

Keyword:
"Inventive step (no)"

Decisions cited:
-

Catchword:
-

Case Number: T 1546/05 - 3.4.03

DECISION
of the Technical Board of Appeal 3.4.03
of 4 June 2008

Appellant:

AT & T Corp.
32 Avenue of the Americas
New York, NY 10013-2412 (US)

Representative:

Kuhnen & Wacker
Patent- und Rechtsanwaltsbüro
Postfach 19 64
D-85319 Freising (DE)

Decision under appeal:

Decision of the Examining Division of the European Patent Office posted 6 July 2005 refusing European application No. 96118549.3 pursuant to Article 97(1) EPC.

Composition of the Board:

Chairman: R. G. O'Connell
Members: G. Eliasson
J. Van Moer

Summary of Facts and Submissions

I. This is an appeal against the refusal of application 96 118 549 for lack of inventive step over

D1: "Voice and facial image integration for person recognition", C. C. Chibelushi et al. in Proceedings of the IEEE International Symposium on Multimedia Technologies and Future Applications, Southampton, UK, 21-23 April 1993, pages 155 to 161; and

D3: DE 44 35 272 A.

II. At oral proceedings before the board, the appellant applicant requested that the decision under appeal be set aside and a patent granted on the basis of the application documents as refused, ie:

claim 1 sent with letter dated 6 May 2005,
claims 2, 3, 5 to 10 as originally filed
claim 4, 11 to 13 sent with letter dated
21 November 2003.

III. Claim 1 reads:

"1. A method for determining authenticity of an individual, said method comprising the steps of:

obtaining audio data (101) of the individual speaking at least one selected phrase;

obtaining video data (101) of the individual speaking said at least one selected phrase;

extracting (103) identifying audio features and video features from said audio data and said video data, respectively;

forming a single feature vector (105) that incorporates said audio features and said video features, said feature vector varying over the duration of the spoken phrase and including the parameters of both the audio and video features that have been extracted;

comparing (107) said feature vector to a stored feature vector of a validated user speaking said at least one selected phrase; and

authenticating (109) said individual if said feature vector and said stored feature vector form a match within a prescribed threshold."

IV. The appellant applicant argued essentially as follows:

(a) In contrast to the method of claim 1, the method of document D1 did not use video features but image data extracted from still images. Furthermore, a *single* feature vector, which incorporates parameters of both the audio and video features, was formed according to the claimed method. As a final step the single feature vector was compared to a *stored* feature vector of a validated user in order to authenticate the user.

(b) Document D3 did not teach anything else than declaring the mere input of the video channel output lines 15 and the audio output lines 25 as providing a combination of a five-component-vector and a 14-component-vector. In the claimed method, a combined, *synchronised* audio-video-vector was formed which was compared with stored vectors in the step of authentication. Neither document D1 nor D3 stored audio-video-vectors of known individuals, since both documents used neural networks in the step of identification. The claimed method had the advantage that it was more accurate than an artificial neural network.

(c) A neural network as disclosed in document D3 had multiple layers including a plurality of units and was used to simulate a predetermined function. The time varying acoustic feature vector and the time varying image feature vector were input to the TDNN 200 in parallel, where weighted summations of the input signals were calculated. Hence the vector generated by incorporating audio and video feature vectors no longer included the intact parameters of both the extracted audio and video features.

Reasons for the Decision

1. The appeal is admissible.

2. *Inventive step*

2.1 Document D1 is considered closest prior art and discloses a method of identifying persons using speech recognition combined with facial images. The image information is extracted from face profile images and processed to extract image features to represent each face. Pre-selected phrases are recorded and extracted in form of cepstra, ie Fourier transforms of the decibel spectrum. The extracted facial and audio features of validated users are modelled in an artificial neural network as a result of a learning process (page 158 "Preliminary investigations"). In the identifying step, the extracted identifying audio and video features of an individual speaking the pre-selected phrases are separately fed into the artificial neural network where a recognition decision is made. In order to improve the reliability of the method, document D1 goes on to propose the use of dynamic facial images in conjunction with the audio features instead of still images of the face (page 157). The proposed method would involve cross-correlation between the motion of visible articulators and the acoustic speech in order to unmask impostors using facial images and voice not originating from the same person (page 160 "New direction").

2.2 The method of claim 1 differs from that of document D1 in that

a) video features are extracted, whereas in the method of document D1, image data from still images are extracted;

b) a *single* feature vector, which incorporates parameters of both the audio and video features, is formed; and

c) the single feature vector is compared to a *stored* feature vector of a validated user in order to authenticate the user, whereas in document D1, the extracted visual and audio features are input into an artificial neural network where a recognition decision is made.

2.3 The above features solve the problem of achieving an accelerated and more accurate classification, or comparison, of the received vector.

2.4 Document D3 discloses a method for speech recognition where visual speech data is used together with audio data. The audio data is extracted using Fourier transforms of the power spectrum from which 14 components are stored (page 9, line 62 to page 10, line 14). The motion of five points of the mouth are extracted (Figure 9) extracted to form a video vector with five components. The extracted audio and video components are combined to one audio-video vector with synchronous components in 1/100 s steps (page 10, lines 32 to 57). A time-delay neural network (TDNN) 200 ("Sprachklassifikator") is used for identifying audio and video features and taking a recognition decision (page 13, lines 34 to 39; Figures 18 and 19).

2.5 Regarding feature a), since document D1 teaches the use of video features in order to enhance the accuracy of the identification and document D3 discloses an implementation of combining audio and video features, the inclusion of feature a) in the method of document D1 would be an obvious measure for the skilled person seeking to improve the method of document D1. Furthermore, as document D3 teaches combining the extracted audio and video features in a combined vector before carrying out the step of identification, the skilled person would also arrive at feature b) by following the teaching of document D3.

2.6 As to feature c), the step of comparing the feature vector to a stored feature vector of a validated user, the board confirms the examining division's finding that the use of vector matching/comparison techniques was one of several straightforward, well-accepted and alternative classification techniques from which the skilled person would select in order to solve the problem posed.

2.7 The appellant applicant argued that a combined audio-visual vector representing a validated user is not stored for comparison in the method of document D3, a fact that was acknowledged in the decision under appeal (see item IV(b) above). The board notes however that the appellant applicant did not contest the examining division's finding that the skilled person would regard the claimed matching/comparison technique as an obvious alternative to the neural network technique disclosed in documents D1 and D3.

2.8 For the above reasons, in the board's judgement, the subject matter of claim 1 does not involve an inventive step within the meaning of Article 56 EPC 1973.

Order

For these reasons it is decided that:

The appeal is dismissed.

Registrar	Chair
S. Sánchez Chiquero	R. G. O'Connell