(19) Europäisches Patentamt
European Patent Office
Office européen des brevets

(11) EP 1 271 469 A1

(12) EUROPEAN PATENT APPLICATION

(43) Date of publication:
02.01.2003 Bulletin 2003/01

(51) Int Cl.⁷: G10L 13/02, G10L 13/08

(21) Application number: 01115216.2

(22) Date of filing: 22.06.2001

(84) Designated Contracting States:
AT BE CH CY DE DK ES FI FR GB GR IE IT LI LU MC NL PT SE TR
Designated Extension States:
AL LT LV MK RO SI

(71) Applicant: Sony International (Europe) GmbH
10785 Berlin (DE)

(72) Inventors:
- Marasek, Krzysztof
Heinrich-Hertz-Str. 1, 70327 Stuttgart (DE)
- Kemp, Thomas
Heinrich-Hertz-Str. 1, 70327 Stuttgart (DE)

(74) Representative: Müller - Hoffmann & Partner
Patentanwälte,
Innere Wiener Strasse 17
81667 München (DE)

(54) Method for generating personality patterns and for synthesizing speech

(57) To mimic the speaking behavior of a given speaker, a method for generating personality patterns in particular for synthesizing speech is proposed in which acoustical as well as non-acoustical speech features (SF) are extracted from a given speech input (SI).

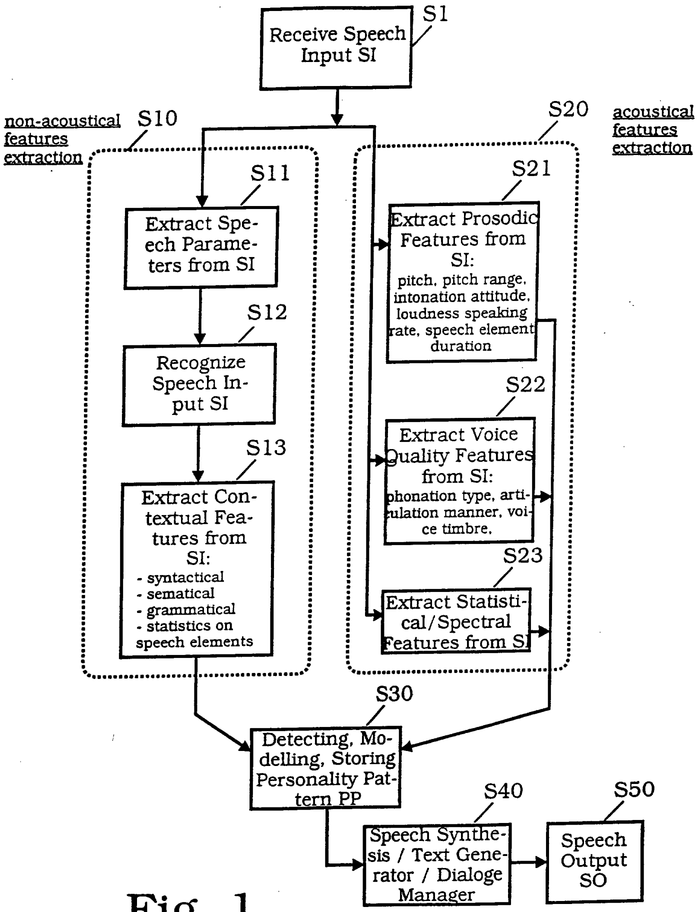
Receive Speech Input SI
S1
non-acoustical features extraction
S10
S11
Extract Speech Parameters from SI
S12
Recognize Speech Input SI
S13
Extract Contextual Features from SI:
- syntactical
- sematical
- grammatical
- statistics on speech elements
S20
acoustical features extraction
S21
Extract Prosodic Features from SI:
pitch, pitch range, intonation attitude, loudness speaking rate, speech element duration
S22
Extract Voice Quality Features from SI:
phonation type, articulation manner, voice timbre.
S23
Extract Statistical/Spectral Features from SI
S30
Detecting, Modelling, Storing Personality Pattern PP
S40
Speech Synthesis / Text Generator / Dialoge Manager
S50
Speech Output SO

Fig. 1

Printed by Jouve, 75001 PARIS (FR)

Description

[0001] The present invention relates to a method for generating personality patterns and to a method for synthesizing speech.

[0002] Nowadays, a large variety of equipment and appliances employ man-machine dialogue systems to ensure an easy and reliable use by a human user. These man-machine dialogue systems are enabled to receive and consider users' utterances, in particular orders and/or inquiries, and to react and respond in an appropriate way. Nevertheless, current speech synthesis systems involved in such man-machine dialogue systems suffer from a lack of personality and naturalness. Although the systems are enabled to deal with the context of the situation in an appropriate way, the prepared and output speech of the dialogue system often sounds monotonically, machine-like, and not embedded into the particular situation.

[0003] It is an object of the present invention to provide a method for generating personality patterns in particular for synthesizing speech and a method for synthesizing speech in which naturalness of the speech and its features can be realized.

[0004] The object is achieved by a method for generating personality patterns, in particular for synthesizing speech, with the features of claim 1. Furtheron, the object is achieved by a method for synthesizing speech according to the characterizing features of claim 11. A system and a computer program product for carrying out the inventive methods are the subject-matter of claims 14 and 15, respectively. Preferred embodiments of the inventive methods are within the scope of the dependent subclaims.

[0005] In the inventive method for generating personality patterns, in particular for synthesizing speech, a speech input is received and/or preprocessed. From the speech input acoustical and/or non-acoustical speech features are extracted. Based on the extracted speech features and/or on models and/or parameters thereof, a personality pattern is generated and/or stored.

[0006] It is therefore a basic idea of the present invention to extract acoustical and alternatively or simultaneously non-acoustical speech features from a received speech input. The speech features are then directly or indirectly used to construct a personality pattern which can lateron be used to reconstruct a speech output with the mimic of the speech input and its speaker. The speech features are therefore parameterized or modeled and included or described in certain models or units.

[0007] According to an embodiment of the inventive method for generating personality patterns, online input speech and/or speech of a speech data base for at least one given speaker are used for receiving said speech input. Using a speech data base enables a system involving the inventive method to generate the personality patterns in advance of an application. That means that, before the system is applied for example in an speech synthesizing unit, a speech model for a single speaker or for a variety of speakers can be constructed. Within the application of the inventive method it is also possible to construct the personality patterns during the application in a speech synthesizing unit in a real time or online manner, so as to adapt a speech output generated in a dialogue system during the application and/or during the dialogue with the user.

[0008] It is an aspect of the present invention to use a large variety of features from the speech input so as to model the personality patterns as good as possible to achieve in an application of a dialogue system a particular natural responding speech output.

[0009] It is therefore an aspect of a further embodiment of the present invention to use prosodic features, voice quality features, global statistic and/or spectral properties, and/or the like as acoustical features.

[0010] Within the class of prosodic features, pitch, pitch range, intonation attitude, loudness, speaking rate, phone duration, speech element duration features, and or the like can be employed.

[0011] Within the class of voice quality features, phonation type, articulation manner, voice timbre features, and/or the like can be employed.

[0012] In the class of non-acoustical features, contextual features and/or the like may be important in accordance to a further advantageous embodiment of the present invention. In particular, syntactical, grammatical, semantical features, and/or the like can be used as contextual features.

[0013] As a human speaker has distinct preferences in constructing sentences, phrases, word combinations, and/or the like, according to a further preferred embodiment of the present invention within the class of non-acoustical features statistical features on the usage, distribution, and/or probability of speech elements - such as words, subword units, syllables, phonemes, phones, and/or the like - and/or combinations of them within said speech input can be used. Additional sentence, phrase, word combination preferences can be evaluated and included into said personality pattern.

[0014] To prepare for the extraction of contextual features or the like, a process of speech recognition is preferably carried out within the inventive method.

[0015] Alternatively or additionally, a process of speaker identification and/or adaptation can be performed, in particular so as to increase the matching rate of the feature extraction and/or of the recognition rate of the process of speech recognition.

[0016] In the inventive method for synthesizing speech, in particular for a man-machine dialogue system, the inventive method for generating personality patterns is employed.

[0017] According to a further embodiment of the inventive method for synthesizing speech, the method for generating personality patterns is essentially carried out in a preprocessing step, in particular based on a speech data base or the like.

[0018] Alternatively or additionally, the method for generating personality patterns can be carried out and/or continued in a continuous, real time, or online manner. This enables a system involving said method for synthesizing speech to adapt its speech output in accordance to the received input during the dialogue.

[0019] Both of the methods for generating personality patterns and/or for synthesizing speech can be configured to create a personality pattern or a speech output which is in some sense complementary to the personality pattern or character assigned to the speaker of the speech input. That means, for instance, that in the case of an emergency call system for activating ambulance or fire alarm services the speaker of the speech input might be excited and/or confused. It might therefore be necessary to calm down the speaking person and this can be achieved by creating a personality pattern for the speech synthesis reflecting a strong and confident and safe character. Additionally, it might also be possible to construct personality patterns for the synthesized speech output which reflects a gender which is complementary to the gender of the speaker of the speech input, i. e. in the case of a male speaker, the system might respond as a female speaker so as to make the dialogue most convenient for the speaking person.

[0020] It is a further aspect of the present invention to provide a system, an apparatus, a device, and/or the like for generating personality patterns and/or for synthesizing speech which is in each case capable of performing and/or realizing the inventive methods for generating personality patterns and/or for synthesizing speech and/or its steps.

[0021] According to a further aspect of the present invention, a computer program product is provided, comprising computer program means which is adapted to perform and/or to realize the inventive method for generating personality patterns and/or for synthesizing speech and/or the steps thereof when it is executed on a computer, a digital signal processing means, and/or the like.

[0022] The aspects of the present invention will become more elucidated taking into account the following remarks:

[0023] After the identification of a speaker, both his relevant voice quality features and his speech itself - as described by any units, such as words, syllables, diphones, sentences, and/or the like - is automatically extracted according to the invention. Also information about preferred sentence structure and word usage are extracted and used to create a speech synthesis system with those characteristics in a completely unsupervised way.

[0024] The starting point for these inventive concepts is the lack of personality of current speech synthesis systems. Prior art systems are developed with text-to-speech (TTS) operation in mind, where intelligibility and naturalness of speech is the most important. For dialogue systems, however, the personality of the dialogue partner is essential, too. Depending on the personality of the artificial dialogue partner, the speaker may be interested in continuation of the dialogue or not. Thus, adding a personality pattern to the speech generated by the device may be crucial for the success of the dialogue device.

[0025] Therefore, it is proposed to collect and store all information about speaking style of the person making conversation with the system or device and to use said information to modify the speaking style of the device.

[0026] The proposed methods can be used to mimic the actual speaker talking to the device but also to equip the device with some different personalities, e. g. gathered from the speaking style of famous people, movie stars, or the like. This can be very attractive for potential customers. The proposed system can be used not only to mimic speaker's behavior but more generally to control the dialogue depending on changing speaking style and emotions of the human partner.

[0027] The collection of features describing the speaker's personality can be done on different levels during the conversation of the human by a dialogue unit. In order to mimic the speaker's voice, the speech signal has to be recorded and segmented into phones, diphones, and/or into other speech units or speech elements in dependence on the speech synthesis method used in the system.

[0028] Prosodic features like pitch, pitch range, attitude of sentence intonation (monotonous or effected), loudness, speaking rate, durations of phones, and/or the like can be collected to characterize the speaker's prosody.

[0029] Voice quality features like phonation type, articulation manner, voice timbre, and/or the like can be automatically extracted from the collected speech data.

[0030] Speaker identification or a speaker identification module are necessary for a proper function of the system.

[0031] The system can also collect all the words recognized from the adherences spoken by the speaker and to generate and evaluate statistics on the usage. This can be used to find the most frequent phrases, words used by a given speaker, and/or the like. Also syntactic information gathered from the recognized phrases can enhance the quality of personality description.

[0032] After all necessary information has been collected, the dialogue system can adjust parameters and units of acoustic output - for example the synthesized waveforms or the like - and modes of text generation to suite the recognized speaker's characteristic.

[0033] The parameterized personality can be stored for future use or can be preprogrammed in the dialogue device. The information can be used to recognize speakers and to change the personality of the system depending on the user's preference or mood, for example in case of a system with a built-in emotion recognition engine.

[0034] The personality can be changed according to the user's wish, preprogrammed sequence or depending on changing speaker's style and emotions of the speaker.

[0035] The main advantage of such a system is the possibility to adapt the dialogue to the given speaker, make the dialogue more attractive, and/or the like. The possibility to mimic certain speakers or to switch between different personalities or speaking styles can be very entertaining and attractive for the user.

[0036] In the following, further advantages and aspects of the present invention will be described taking reference to the accompanying figure.

Fig. 1 is a schematical block diagram describing a preferred embodiment of a method for synthesizing speech employing an embodiment of the inventive method for generating personality patterns.

[0037] The schematical block diagram of Fig. 1 shows a preferred embodiment of the inventive method for a synthesizing speech employing an embodiment of the inventive method for generating personality pattern from a given received speech input SI.

[0038] In step S1, speech input S1 is received. In a first section S10 of the inventive method for synthesizing speech, non-acoustic features are extracted from the received speech input SI. In a second section S20 of the inventive method for synthesizing speech, acoustical features are extracted from the received speech input SI. The sections S10 and S20 can be performed paralely or sequentially on a given device or apparatus.

[0039] In the first section S10 for extracting non-acoustical features from the speech input S1 in a first step S11, speech parameters are extracted from said speech input SI. In a second step S12, the speech input S1 is fed into a speech recognizer to analyze the content and the context of the received speech input SI.

[0040] Based on the recognition result, in a following step S13 contextual features are extracted from said speech input S1, in particular syntactical, semantical, grammatical, and statistical information on particular speech elements are obtained.

[0041] In the embodiment of Fig. 1, the second section S20 of the inventive method for synthesizing speech consists of three steps S21, S22, and S23 to be performed independently from each other.

[0042] In the first step S21 of the second section S20 for extracting acoustical features, prosodic features are extracted from the received speech input SI. Said prosodic feature may comprise features of pitch, pitch range, intonation attitude, loudness, speaking rate, speech element duration, and/or the like.

[0043] In a second step S22, voice quality features are extracted from the given received speech input SI, for instance phonation type, articulation manner, voice timbre features, and/or the like.

[0044] Finally, in a third and final step S23 of the second section S20, statistical/spectral features are extracted from the given speech input SI.

[0045] The non-acoustical features and the acoustical features obtained from sections S10 and S20 are merged in a following postprocessing step S30 to detect, model, and store a personality pattern PP for the given speaker.

[0046] The data describing the personality pattern PP for the current speaker are fed into a following step S40 which includes the steps of speech synthesis, text generation, and dialogue managing from which a responsive speech output SO is generated and then output in a final step S50.

Claims

1. Method for generating personality patterns, in particular for synthesizing speech, wherein:

   - speech input (SI) is received and/or preprocessed,
   - acoustical and/or non-acoustical speech features (SF) are extracted from said speech input (SI),
   - based on the extracted speech features (SF) or on models or parameters thereof a personality pattern (PP) is generated and/or stored.

2. Method according to claim 1, wherein online input speech and/or speech of a speech data base for at least one given speaker are used for receiving said speech input (SI).

3. Method according to anyone of the preceding claims, wherein prosodic features, voice quality features, global statistical, and/or spectral properties, and/or the like are used as acoustical features.

4. Method according to claim 3, wherein pitch, pitch range, intonation attitude, loudness, speaking rate, phone duration, speech element duration features, and/or the like are used as prosodic features.

5. Method according to anyone of the claims 3 or 4, wherein phonation type, articulation manner, voice timbre features, and/or the like are used as voice quality features.

6. Method according to anyone of the preceding claims, wherein contextual features, and/or the like are used as said non-acoustical features.

7. Method according to claim 6, wherein syntactical, grammatical, semantical features, and/or the like are used as contextual features.

8. Method according to anyone of the claims 6 or 7, wherein statistical features on the usage, distribution, and/or probability of speech elements - such as words, subword units, syllables, phonemes, phones, and/or the like - and/or combinations of them within said speech input (SI) are used as non-acoustical features.

9. Method according to anyone of the preceding claims, wherein a process of speech recognition is carried out, in particular to prepare the extraction of contextual features and/or the like.

10. Method according to anyone of the preceding claims, wherein a process of speaker identification and/or adaptation is performed, in particular so as to increase the matching rate of the feature extraction and/or of the recognition rate of the process of speech recognition.

11. Method for synthesizing speech, in particular for a man-machine dialogue system, wherein the method for generating personality patterns according to anyone of the claims 1 to 10 is employed.

12. Method according to claim 11, wherein the method for generating personality patterns is essentially carried out in a preprocessing step, in particular based on a speech data base or the like.

13. Method according to anyone of the claims 11 or 12, wherein the method for generating personality patterns is carried out and/or continued in a continuous, real time, or online manner.

14. System for generating personality patterns and/or for synthesizing speech which is capable of performing and/or realizing the method for generating personality patterns according to anyone of the claims 1 to 10 and/or the method for synthesizing speech according to anyone of the claims 11 to 13 and/or the steps thereof.

15. Computer program product, comprising computer program means adapted to perform and/or to realize the method for generating personality patterns according to anyone of the claims 1 to 10 and/or the method for synthesizing speech according to anyone of the claims 11 to 13 and/or the steps thereof when it is executed on a computer, a digital signal processing means, and/or the like.

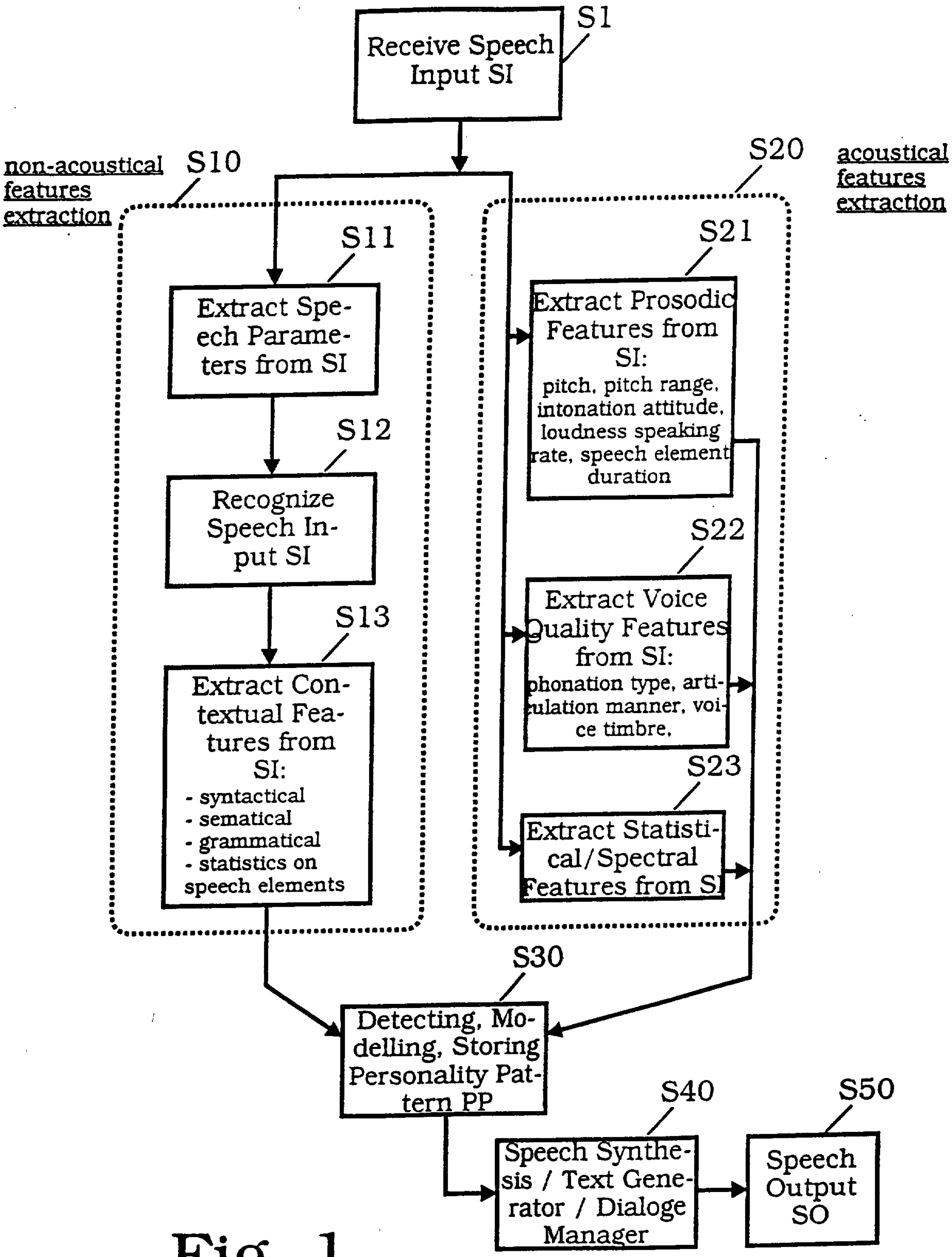
Receive Speech Input SI
S1
non-acoustical features extraction
S10
S20
acoustical features extraction
S11
Extract Speech Parameters from SI
S12
Recognize Speech Input SI
S13
Extract Contextual Features from SI:
- syntactical
- sematical
- grammatical
- statistics on speech elements
S21
Extract Prosodic Features from SI:
pitch, pitch range, intonation attitude, loudness speaking rate, speech element duration
S22
Extract Voice Quality Features from SI:
phonation type, articulation manner, voice timbre,
S23
Extract Statistical/Spectral Features from SI
S30
Detecting, Modelling, Storing Personality Pattern PP
S40
Speech Synthesis / Text Generator / Dialoge Manager
S50
Speech Output SO

Fig. 1

European Patent Office

EUROPEAN SEARCH REPORT

Application Number
EP 01 11 5216

DOCUMENTS CONSIDERED TO BE RELEVANT

Category	Citation of document with indication, where appropriate, of relevant passages	Relevant to claim	CLASSIFICATION OF THE APPLICATION (Int.Cl.7)
X	US 5 278 943 A (GASPER ELON ET AL) 11 January 1994 (1994-01-11)	1,2,11, 14,15	G10L13/02 G10L13/08
Y	* abstract * * column 4, line 53 - column 6, line 9 * * column 13, line 14 - column 48, line 2 * ---	3-5	
Y	JANET E. CAHN: "The Generation of Affect in Synthesized Speech" JOURNAL OF THE AMERICAN VOICE I/O SOCIETY, ‘Online! vol. 8, July 1990 (1990-07), pages 1-19, XP002183399 Retrieved from the Internet: <URL:http://www.media.mit.edu/{cahn/masters-thesis.htm> ‘retrieved on 2001-11-20! * page 3 - page 6 * ---	3-5	
Y	KLASMEYER ET AL: "The perceptual importance of selected voice quality parameters" ACOUSTICS, SPEECH, AND SIGNAL PROCESSING, 1997. ICASSP-97., 1997 IEEE INTERNATIONAL CONFERENCE ON MUNICH, GERMANY 21-24 APRIL 1997, LOS ALAMITOS, CA, USA,IEEE COMPUT. SOC, US, 21 April 1997 (1997-04-21), pages 1615-1618, XP010226301 ISBN: 0-8186-7919-0 * abstract * ---	4	TECHNICAL FIELDS SEARCHED (Int.Cl.7) G10L
A	WO 99 12324 A (BACK WILLIAM K ;HOLLINS JACK (US)) 11 March 1999 (1999-03-11) * page 1, line 23 - line 25; claim 1; figure 2A * --- -/--	1,11,14, 15	

The present search report has been drawn up for all claims

Place of search	Date of completion of the search	Examiner
MUNICH	20 November 2001	De Vos, L

CATEGORY OF CITED DOCUMENTS

X : particularly relevant if taken alone
Y : particularly relevant if combined with another document of the same category
A : technological background
O : non-written disclosure
P : intermediate document

T : theory or principle underlying the invention
E : earlier patent document, but published on, or after the filing date
D : document cited in the application
L : document cited for other reasons

& : member of the same patent family, corresponding document

European Patent Office

EUROPEAN SEARCH REPORT

Application Number
EP 01 11 5216

DOCUMENTS CONSIDERED TO BE RELEVANT

Category	Citation of document with indication, where appropriate, of relevant passages	Relevant to claim	CLASSIFICATION OF THE APPLICATION (Int.Cl.7)
A	US 6 144 938 A (ALBERT ROY D ET AL) 7 November 2000 (2000-11-07) * abstract; claims 9,22-25; figures 15,19 * * column 20, line 58 - column 21, line 4 * -----	1,11,14, 15	
			TECHNICAL FIELDS SEARCHED (Int.Cl.7)

The present search report has been drawn up for all claims

Place of search	Date of completion of the search	Examiner
MUNICH	20 November 2001	De Vos, L

CATEGORY OF CITED DOCUMENTS

X : particularly relevant if taken alone
Y : particularly relevant if combined with another document of the same category
A : technological background
O : non-written disclosure
P : intermediate document

T : theory or principle underlying the invention
E : earlier patent document, but published on, or after the filing date
D : document cited in the application
L : document cited for other reasons

& : member of the same patent family, corresponding document

ANNEX TO THE EUROPEAN SEARCH REPORT ON EUROPEAN PATENT APPLICATION NO.

EP 01 11 5216

This annex lists the patent family members relating to the patent documents cited in the above-mentioned European search report. The members are as contained in the European Patent Office EDP file on
The European Patent Office is in no way liable for these particulars which are merely given for the purpose of information.

20-11-2001

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
US 5278943 A	11-01-1994	NONE	
WO 9912324 A	11-03-1999	US 6317486 B1 WO 9912324 A1	13-11-2001 11-03-1999
US 6144938 A	07-11-2000	EP 1074017 A1 WO 9957714 A1	07-02-2001 11-11-1999

For more details about this annex : see Official Journal of the European Patent Office, No. 12/82