(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
3 January 2014 (03.01.2014)

WIPO | PCT

(10) International Publication Number
WO 2014/003748 A1

(51) International Patent Classification:
G10L 15/20 (2006.01) *G06F 17/30* (2006.01)
G10L 15/24 (2013.01)

(21) International Application Number:
PCT/US2012/044514

(22) International Filing Date:
28 June 2012 (28.06.2012)

(25) Filing Language: English

(26) Publication Language: English

(71) Applicant *(for all designated States except US)*: **NUANCE COMMUNICATIONS, INC.** [US/US]; 1 Wayside Road, Suite 100, Burlington, MA 01803-4613 (US).

(72) Inventors; and
(75) Inventors/Applicants *(for US only)*: **WILLETT, Daniel** [DE/DE]; Haselnussgasse 9, 65396 Walluf (DE). **LÖÖF, Karl, Jonas** [SE/DE]; Franzstr. 95, 52064 Aachen (DE). **PAN, Yue** [CN/US]; 46 Charles Road, Winchester, MA 01890 (US). **PINTO, Joel** [IN/DE]; Schlossparkstrasse 6, 52072 Aachen (DE). **GOLLAN, Christian** [DE/DE]; Augustastr. 18, 52070 Aachen (DE).

(74) Agents: **SUNSTEIN, Bruce D.** et al.; Sunstein Kann Murphy & Timbers LLP, 125 Summer Street, Boston, MA 02110 (US).

(81) Designated States *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IS, JP, KE, KG, KM, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LT, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

(84) Designated States *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Published:

— *with international search report (Art. 21(3))*

(54) **Title**: META-DATA INPUTS TO FRONT END PROCESSING FOR AUTOMATIC SPEECH RECOGNITION

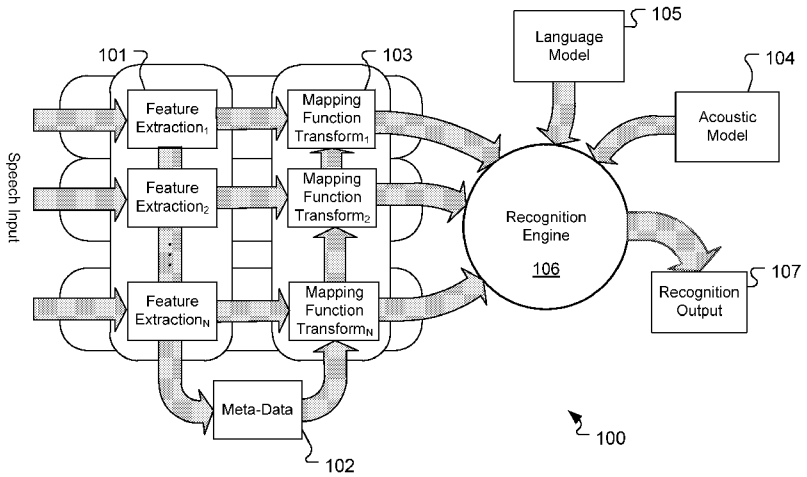


Fig. 1

(57) **Abstract**: A computer-implemented method is described for front end speech processing for automatic speech recognition. A sequence of speech features which characterize an unknown speech input provided on an audio input channel and associated meta-data which characterize the audio input channel are received. The speech features are transformed with a computer process that uses a trained mapping function controlled by the meta-data, and automatic speech recognition is performed of the transformed speech features.

WO 2014/003748 A1

Title

META-DATA INPUTS TO FRONT END PROCESSING FOR AUTOMATIC SPEECH RECOGNITION

Field of the Invention

[0001] The invention generally relates to automatic speech recognition (ASR), and more specifically, to use of met-data in front end feature processing for ASR.

Background Art

[0002] An automatic speech recognition (ASR) system determines a semantic meaning of a speech input. Typically, the input speech is processed into a sequence of digital speech feature frames. Each speech feature frame can be thought of as a multi-dimensional vector that represents various characteristics of the speech signal present during a short time window of the speech. For example, the multi-dimensional vector of each speech frame can be derived from cepstral features of the short time Fourier transform spectrum of the speech signal (MFCCs)—the short time power or component of a given frequency band—as well as the corresponding first- and second-order derivatives ("deltas" and "delta-deltas"). In a continuous recognition system, variable numbers of speech frames are organized as "utterances" representing a period of speech followed by a pause, which in real life loosely corresponds to a spoken sentence or phrase.

[0003] The ASR system compares the input speech frames to find statistical models that best match the speech feature characteristics and then determines a corresponding representative text or semantic meaning associated with the statistical models. Modern statistical models are state sequence models such as hidden Markov models (HMMs) that model speech sounds (usually phonemes) using mixtures of Gaussian distributions. Often these statistical models represent phonemes in specific contexts, referred to as PELs (Phonetic Elements), e.g. triphones or phonemes with known left and/or right contexts. State sequence models can be scaled up to represent words as connected sequences of acoustically modeled phonemes, and phrases or sentences as connected sequences of words. When the statistical models are organized together as words, phrases, and sentences, additional language-related information is also typically incorporated into the models in the form of language modeling.

[0004] The words or phrases associated with the best matching model structures are referred to as recognition candidates or hypotheses. A system may produce a single best recognition candidate – the recognition result – or a list of several hypotheses, referred to as an N-best list. Further details regarding continuous speech recognition are provided in U.S. Patent No. 5,794,189, entitled "Continuous Speech Recognition," and U.S. Patent No. 6,167,377, entitled "Speech Recognition Language Models," the contents of which are incorporated herein by reference.

[0005] In many ASR applications—for example, cloud-based client-server ASR—the speech recognizer is exposed to speech data acquired from many different devices operating in various acoustic environments and from different applications such as messaging or voice search. Device type, microphone type (and position on the device) as well as acoustic environment have an influence on the observed audio. To a somewhat lesser degree, application type has this effect as it affects speaking style and the way users generally hold and operate the device. Furthermore, the acoustic signal as observed at the ASR system usually has passed through a data channel of limited bandwidth that requires application of an encoding that comes along with information loss (lossy codec). Often, this information on device- and microphone-type, on application, on codec setting is available to the server recognizer as meta-data that characterizes such aspects of the audio input channel.

[0006] Meta-data typically is categorical information that can attain one of a finite set of values. Representing this data as input for neural networks or other forms of linear and non-linear transformations allows many approaches. Categorical information can be represented as numerical values, for example, by representing each category as an integer or using 1-of-N encoding.

Summary of the Invention

[0007] Embodiments of the present invention include a computer-implemented method employing at least one hardware implemented computer processor for front end speech processing for automatic speech recognition. A sequence of speech features which

characterize an unknown speech input provided on an audio input channel and associated meta-data which characterize the audio input channel are received. The speech features are transformed with a computer process that uses a trained mapping function controlled by the meta-data. A computer process is then used to perform automatic speech recognition of the transformed speech features.

[0008] The meta-data may include meta-data characterizing a source channel which generates the unknown speech input, an audio codec applied when generating the speech features, a source environment in which the unknown speech input is generated, and/or a source application which generates the unknown speech input. The mapping function may be, for example, a multilayered perceptron (MLP) mapping function, a deep belief network (DBN) mapping function, and/or a feature-space maximum mutual information (fMMI) mapping function. The steps may be performed in real time with minimum response latency.

Brief Description of the Drawings

[0009] Figure 1 shows an example of ASR system according to one embodiment of the present invention.

[0010] Figure 2 shows various logical steps in a method of front end speech feature processing using meta-data according to an embodiment of the present invention.

Detailed Description of Specific Embodiments

[0011] Various embodiments of the present invention directly use the meta-data commonly included in the audio input channel signal for front-end feature processing in ASR. This can provide better channel-, environment- and application-normalized speech features and helps address the impact of the channel, environment and application on the observed speech audio in server-side cloud-based ASR to improve system performance (general accuracy, compute speed and latency) as well as robustness to outlier channel and environment types.

[0012] The direct exploitation of the meta-data in the speech feature front-end can come in various specific arrangements. For example, the meta-data can be used as an additional input to trained mapping functions such as a Multi-Layer-Perceptron (MLP) or Deep Belief Network (DBN) used as a feature generator. For discriminative feature spaces such as feature-space maximum mutual information (fMMI) that utilize a set of Gaussian distributions as generators of a new high dimensional feature space, meta-data can be integrated using a meta-data enhanced mapping prior to input to the set of Gaussians, e.g., an MLP applied as a feature generator of what goes into the fMMI transformation. Alternatively, meta-data integration into fMMI can be approached via replacement of the set of Gaussians via Neural Network-based posterior estimates that see the meta-data as input.

[0013] Figure 1 shows an example of ASR system and Figure 2 shows various logical steps in a method of front end speech feature processing using meta-data according to an embodiment of the present invention. An unknown speech signal from one of a variety of source channels 1-N is input to a feature extraction module **101** of the ASR system **100.** The feature extraction module **101** uses a computer process to create pre-processed audio that includes a sequence of representative speech features which characterize the input signal and meta-data **102** which characterize the audio input channel, step **201**.

[0014] The meta-data **102** may include meta-data characterizing a source channel which generates the unknown speech input, a source environment in which the unknown speech input is generated, and/or a source application which generates the unknown speech input.

[0015] One specific example of a meta-data based additional input to a feature transformation would be a case where microphone type information is available as meta-data, and assuming the categories:

a) built-in mic
b) external mic

This information could be fed into the feature transformation as a binary input assuming the values 0 and 1. The size of the data stream after audio encoding (codec) is one specific example of meta-data that directly comes as a number. Depending on the codec setting,

this can be characterized as arbitrary numbers of bits per second, e.g., 28.8 kbps. A more complex example might be a device type that has more categories, like a hypothetical:

	iPhn4S	iPad2	iOS	headset
a) iPhone4S	1	0	1	1
b) other iPhone	0	0	1	1
c) iPad2	0	1	1	0
d) other iPad	0	0	1	0
e) other headset	0	0	0	1
f) other tablet	0	0	0	0

Then the meta-data information could be decoded as a string of binary values.

[0016] A trained mapping function transform **103** controlled by the meta-data **102** then transform the pre-processed audio speech features, step **202**, to produce a sequence of transformed speech features. The transformation generally provides dimensionality reduction, a better fit of the speech features to the acoustic model, and normalization (reduction of channel-, speaker-, and/or environment-impact) which is extended by the direct input of the meta-data information.

[0017] For example, the trained mapping function transform **103** may be a segmentation multilayered perceptron (MLP) that performs a mapping function trained on segmentation for the input speech features and meta-data to produce monophone output states (e.g., ~150). The output states of the segmentation MLP may then be transformed by a feature-space maximum mutual information (fMMI) mapping function trained on the MLP-output posteriors rather than conventional Gaussian posteriors. *See* Hermansky et al., *Tandem Connectionist Feature Extraction for Conventional HMM Systems*, in Proc. ICASSP, 2000; Daniel Povey, *Improvements to fMPE for Discriminative Training of Features*, Interspeech 2005; which are incorporated herein by reference. Other embodiments could use other specific arrangements, for example, using a deep belief network (DBN), cepstral mean normalization (CMN), cepstral variance normalization (CVN), linear discriminant analysis (LDA), and/or feature-space maximum likelihood linear regression (fMLLR), etc. This front-end feature processing may be performed in real time with minimum response latency for cloud-based ASR as for a dictation application. Speech recognition engine **106**

then compares the transformed speech features to acoustic models **104** and language model **105** to determine a recognition output **107**, step **203**.

[0018] Embodiments of the invention may be implemented in any conventional computer programming language. For example, preferred embodiments may be implemented in a procedural programming language (*e.g.*, "C") or an object oriented programming language (*e.g.*, "C++", Python). Alternative embodiments of the invention may be implemented as pre-programmed hardware elements, other related components, or as a combination of hardware and software components.

[0019] Embodiments can be implemented as a computer program product for use with a computer system, for example, as a series of computer instructions fixed either on a tangible medium, such as a computer readable medium (*e.g.,* a diskette, CD-ROM, ROM, or fixed disk) or transmittable to a computer system, via a modem or other interface device, such as a communications adapter connected to a network over a medium. The medium may be either a tangible medium (*e.g.,* optical or analog communications lines) or a medium implemented with wireless techniques (*e.g.,* microwave, infrared or other transmission techniques). The series of computer instructions embodies all or part of the functionality previously described herein with respect to the system. Those skilled in the art should appreciate that such computer instructions can be written in a number of programming languages for use with many computer architectures or operating systems. Furthermore, such instructions may be stored in any memory device, such as semiconductor, magnetic, optical or other memory devices, and may be transmitted using any communications technology, such as optical, infrared, microwave, or other transmission technologies. It is expected that such a computer program product may be distributed as a removable medium with accompanying printed or electronic documentation (*e.g.,* shrink wrapped software), preloaded with a computer system (*e.g.,* on system ROM or fixed disk), or distributed from a server or electronic bulletin board over the network (*e.g.,* the Internet). Of course, some embodiments of the invention may be implemented as a combination of both software (*e.g.*, a computer program product) and hardware. Still other embodiments of the invention are implemented as entirely hardware, or entirely software (*e.g.*, a computer program product).

[0020] Although various exemplary embodiments of the invention have been disclosed, it should be apparent to those skilled in the art that various changes and modifications can be made which will achieve some of the advantages of the invention without departing from the true scope of the invention.

What is claimed is:

1. A computer-implemented method employing at least one hardware implemented computer processor for front end speech processing for automatic speech recognition, the method comprising:

receiving with a computer process:

i. a sequence of speech features which characterize an unknown speech input provided on an audio input channel, and

ii. associated meta-data which characterize the input audio channel;

transforming the speech features with a computer process using a trained mapping function controlled by the meta-data; and

using a computer process to perform automatic speech recognition of the transformed speech features.

2. A method according to claim 1, wherein the meta-data includes meta-data characterizing a source channel which generates the unknown speech input.

3. A method according to claim 1, wherein the meta-data includes meta-data characterizing a source environment in which the unknown speech input is generated.

4. A method according to claim 1, wherein the meta-data includes meta-data characterizing a source application which generates the unknown speech input.

5. A method according to claim 1, wherein the meta-data includes meta-data characterizing an audio codec applied when generating the speech features.

6. A method according to claim 1, wherein the trained mapping function includes a multilayered perceptron (MLP) mapping function.

7. A method according to claim 1, wherein the trained mapping function includes a deep belief network (DBN) mapping function.

8. A method according to claim 1, wherein the trained mapping function includes a

feature-space maximum mutual information (fMMI) mapping function.

9. A method according to claim 1, wherein the steps are performed in real time with minimum response latency.

10. A computer program product encoded in a non-transitory computer-readable medium for front end speech processing for automatic speech recognition, the product comprising:

program code for receiving :

i. a sequence of speech features which characterize an unknown speech input provided on an audio input channel, and

ii. associated meta-data which characterize the input audio channel;

program code for transforming the speech features using a trained mapping function controlled by the meta-data; and

program code for performing automatic speech recognition of the transformed speech features.

11. A product according to claim 10, wherein the meta-data includes meta-data characterizing a source channel which generates the unknown speech input.

12. A product according to claim 10, wherein the meta-data includes meta-data characterizing a source environment in which the unknown speech input is generated.

13. A product according to claim 10, wherein the meta-data includes meta-data characterizing a source application which generates the unknown speech input.

14. A product according to claim 10, wherein the meta-data includes meta-data characterizing an audio codec applied when generating the speech features.

15. A product according to claim 10, wherein the trained mapping function includes a multilayered perceptron (MLP) mapping function.

16. A product according to claim 10, wherein the trained mapping function includes a deep belief network (DBN) mapping function.

17. A product according to claim 10, wherein the trained mapping function includes a feature-space maximum mutual information (fMMI) mapping function.

18. A product according to claim 10, wherein the program code is executed in real time with minimum response latency.

Applicants: Nuance Communications, Inc. et al.
Title: Meta-Data Inputs to Front End Processing for Automatic Speech Recognition
Application No.: Not yet assigned
Filing Date: June 28, 2012
Docket No.: 2639/C81WO

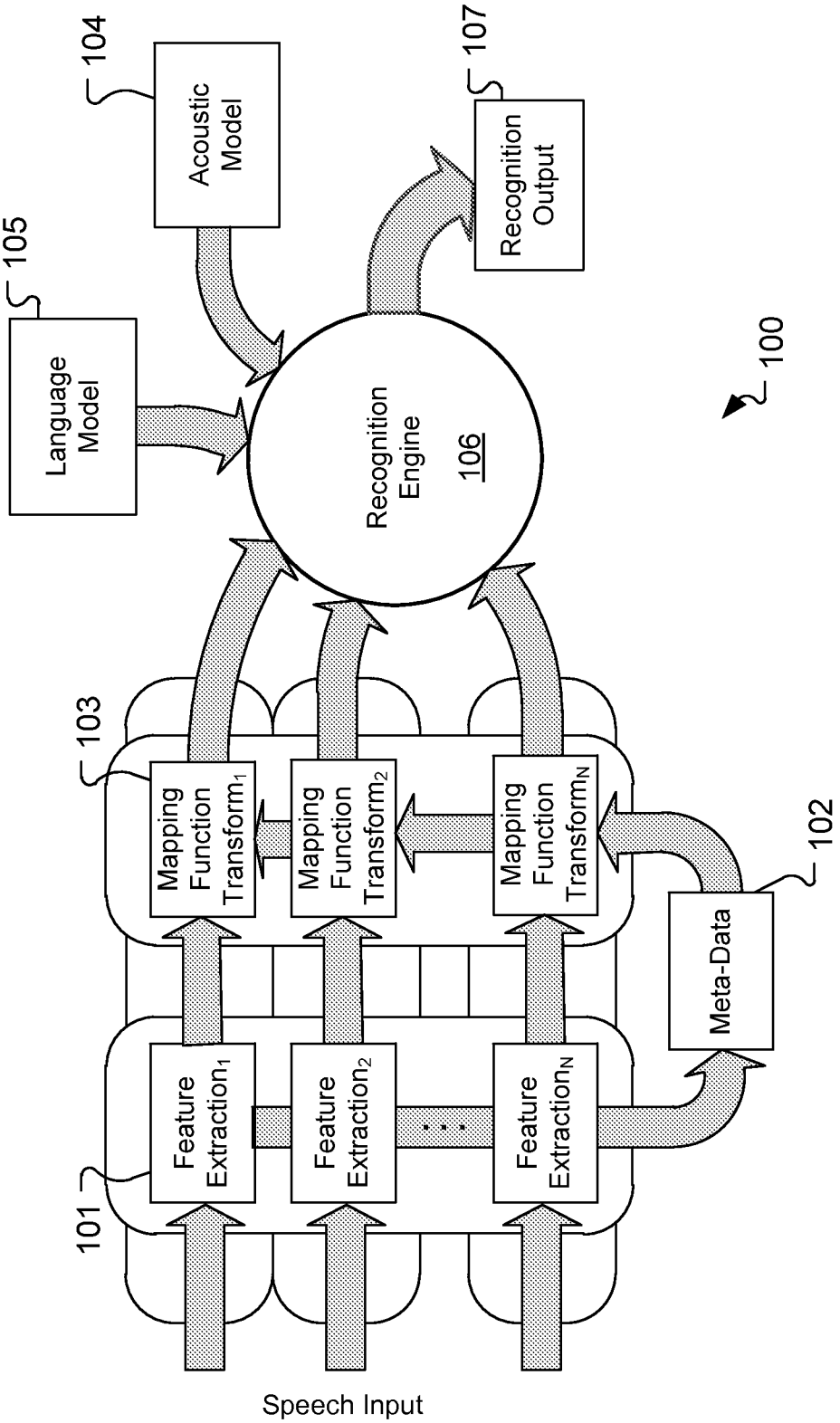


Fig. 1

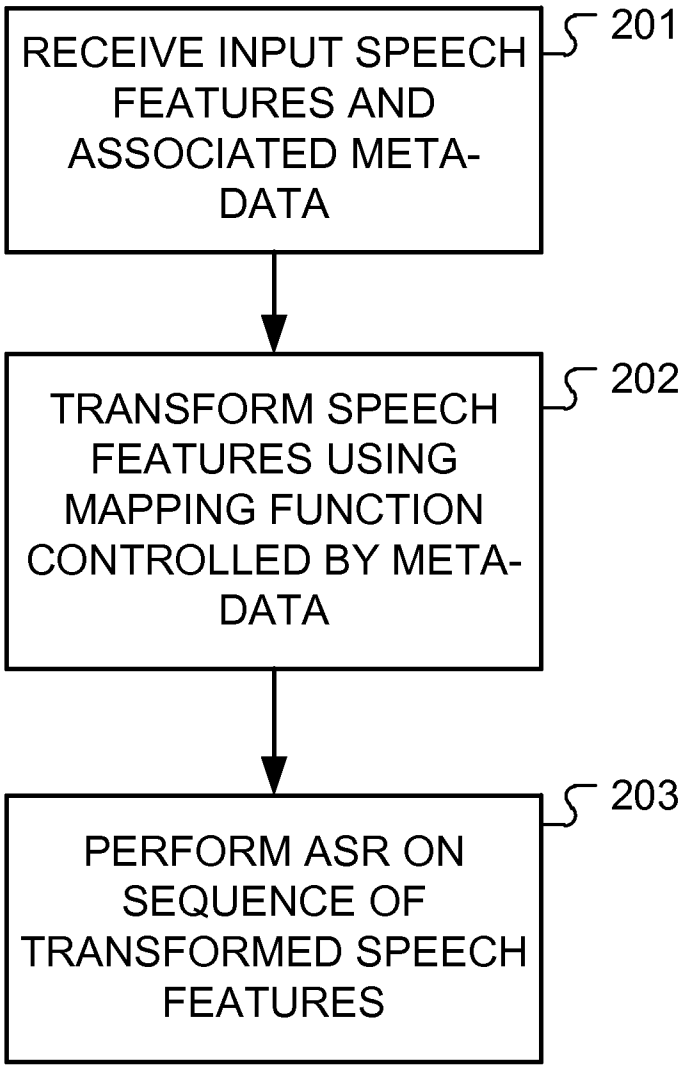

RECEIVE INPUT SPEECH FEATURES AND ASSOCIATED META-DATA
201
TRANSFORM SPEECH FEATURES USING MAPPING FUNCTION CONTROLLED BY META-DATA
202
PERFORM ASR ON SEQUENCE OF TRANSFORMED SPEECH FEATURES
203

Fig. 2

INTERNATIONAL SEARCH REPORT

International application No
PCT/US2012/044514

A. CLASSIFICATION OF SUBJECT MATTER
INV. G10L15/20 G10L15/24 G06F17/30
ADD.

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
G10L G06F

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

EPO-Internal, WPI Data

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X Y	US 5 737 485 A (FLANAGAN JAMES L [US] ET AL) 7 April 1998 (1998-04-07) figure 1 column 1, line 13 - column 2, line 50 column 3, line 47 - column 4, line 41 column 5, lines 36-39 column 9, lines 45-47 -----	1-6, 9-15,18 7,8,16, 17
Y	ABDEL-RAHMAN MOHAMED ET AL: "Deep Belief Networks using discriminative features for phone recognition", ACOUSTICS, SPEECH AND SIGNAL PROCESSING (ICASSP), 2011 IEEE INTERNATIONAL CONFERENCE ON, IEEE, 22 May 2011 (2011-05-22), pages 5060-5063, XP032001819, DOI: 10.1109/ICASSP.2011.5947494 ISBN: 978-1-4577-0538-0 sections 1-3 -----	7,8,16, 17

☐ Further documents are listed in the continuation of Box C.

☒ See patent family annex.

* Special categories of cited documents :

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier application or patent but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
25 September 2012	05/10/2012

Name and mailing address of the ISA/	Authorized officer
European Patent Office, P.B. 5818 Patentlaan 2 NL - 2280 HV Rijswijk Tel. (+31-70) 340-2040, Fax: (+31-70) 340-3016	Tilp, Jan

1

INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No
PCT/US2012/044514

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
US 5737485 A	07-04-1998	NONE	