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

(11) **EP 2 959 475 B1**

(12) **EUROPEAN PATENT SPECIFICATION**

(45) Date of publication and mention of the grant of the patent:
08.02.2017 Bulletin 2017/06

(21) Application number: **13731759.0**

(22) Date of filing: **26.06.2013**

(51) Int Cl.:
G10L 15/14 *(2006.01)*

(86) International application number:
PCT/EP2013/063330

(87) International publication number:
WO 2014/177232 (06.11.2014 Gazette 2014/45)

(54) A SPEECH RECOGNITION SYSTEM AND A METHOD OF USING DYNAMIC BAYESIAN NETWORK MODELS

SPRACHERKENNUNGSSYSTEM UND VERFAHREN ZUR VERWENDUNG DYNAMISCHER BAYES-NETZWERKMODELLE

SYSTÈME DE RECONNAISSANCE DE LA PAROLE ET PROCÉDÉ D'UTILISATION DE MODÈLES DE RÉSEAU DE BAYES DYNAMIQUE

(84) Designated Contracting States:
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

(30) Priority: **01.05.2013 PL 40372413**

(43) Date of publication of application:
30.12.2015 Bulletin 2015/53

(73) Proprietor: **Akademia Gorniczo-Hutnicza im. Stanislawa
Staszica w Krakowie
30-059 Krakow (PL)**

(72) Inventors:
- **ZIÓLKO, Bartosz
PL-30-364 Kraków (PL)**
- **JADCZYK, Tomasz
PL-42-605 Tarnowskie Góry (PL)**

(74) Representative: **Eupatent.pl
ul. Zeligowskiego 3/5
90-752 Lodz (PL)**

(56) References cited:
US-B1- 6 292 776

- **TODD A STEPHENSON HERVE BOURLARD SAMY BENGIO ANDREW C MORRIS DALLE MOLLE INSTITUTE FOR PERCEPTUAL ARTIFICIAL INTELLIGENCE (IDIAP): "AUTOMATIC SPEECH RECOGNITION USING DYNAMIC BAYESIAN NETWORKS WITH BOTH ACOUSTIC AND ARTICULATORY VARIABLES", 6TH. INTERNATIONAL CONFERENCE OF SPOKEN LANGUAGE PROCESSING. ICSLP 2000. BEIJING, CHINA, OCT. 16 - 20, 2000; [INTERNATIONAL CONFERENCE OF SPOKEN LANGUAGE PROCESSING], CENTER FOR SPOKEN LANGUAGE RESEARCH, 16 October 2000 (2000-10-16), XP007010757, ISBN: 978-7-80150-144-8**
- **T.J. HAZEN: "Visual model structures and synchrony constraints for audio-visual speech recognition", IEEE TRANSACTIONS ON AUDIO, SPEECH AND LANGUAGE PROCESSING, vol. 14, no. 3, 1 May 2006 (2006-05-01), pages 1082-1089, XP055112509, ISSN: 1558-7916, DOI: 10.1109/TSA.2005.857572**
- **GOWDY J ET AL: "DBN based multi-stream models for audio-visual speech recognition", ACOUSTICS, SPEECH, AND SIGNAL PROCESSING, 2004. PROCEEDINGS. (ICASSP '04). IEEE INTERNATIONAL CONFERENCE ON MONTREAL, QUEBEC, CANADA 17-21 MAY 2004, PISCATAWAY, NJ, USA,IEEE, PISCATAWAY, NJ, USA, vol. 1, 17 May 2004 (2004-05-17), pages 993-996, XP010717798, DOI: 10.1109/ICASSP.2004.1326155 ISBN: 978-0-7803-8484-2**

Printed by Jouve, 75001 PARIS (FR)

(Cont. next page)

- XAVIER DOMONT ET AL: "Hierarchical spectro-temporal features for robust speech recognition", ACOUSTICS, SPEECH AND SIGNAL PROCESSING, 2008. ICASSP 2008. IEEE INTERNATIONAL CONFERENCE ON, IEEE, PISCATAWAY, NJ, USA, 31 March 2008 (2008-03-31), pages 4417-4420, XP031251577, ISBN: 978-1-4244-1483-3
- KATE SAENKO ET AL: "AN ASYNCHRONOUS DBN FOR AUDIO-VISUAL SPEECH RECOGNITION", SPOKEN LANGUAGE TECHNOLOGY WORKSHOP, 2006. IEEE, IEEE, PI, 1 December 2006 (2006-12-01), pages 154-157, XP031056424, ISBN: 978-1-4244-0872-6
- STÉPHANE DUPONT ET AL: "Audio-Visual Speech Modeling for Continuous Speech Recognition", IEEE TRANSACTIONS ON MULTIMEDIA, IEEE SERVICE CENTER, PISCATAWAY, NJ, US, vol. 2, no. 3, 1 September 2000 (2000-09-01), XP011036218, ISSN: 1520-9210
- ASTRID HAGEN ET AL: "USING MULTIPLE TIME SCALES IN THE FRAMEWORK OF MULTI-STREAM SPEECH RECOGNITION", INTERNATIONAL CONF. ON SPOKEN LANGUAGE PROCESSING, BEIJING 2000, 16 October 2000 (2000-10-16), XP007011060,

Description

TECHNICAL FIELD

[0001] The object of the present invention is a speech recognition system and a method of using Bayesian networks for this purpose. In particular, such an automatic speech recognition system that is applicable in dialog systems for advertising and for informational purposes. Implementations of dialog systems may take a form of information kiosks or booths that will begin a conversation with a customer or viewer and will present appropriate multimedia content.

BACKGROUND ART

[0002] Speech recognition systems are becoming more and more common in everyday life. For example they have been implemented in information call centers such as for public transport. These systems are, however, still frequently operated by keypads and text as a source of input information, instead of speech.

[0003] There are known various kinds of computerized interactive kiosks allowing for conducting a conversation with a user. For example, a US patent US6256046 discloses an active public user interface in a computerized kiosk sensing persons by processing of visual data, by using movement and color analysis to detect changes in the environment indicating the presence of people. Interaction spaces are defined and the system records an initial model of its environment which is updated over time to reflect the addition or subtraction of inanimate objects and to compensate for lighting changes. The system develops models of the moving objects and is thereby able to track people as they move about the interaction spaces. A stereo camera system further enhances the system's ability to sense location and movement. The kiosk presents audio and visual feedback in response to what it "sees".

[0004] A US patent application US20080204450 discloses a system, method and program product for providing a virtual universe in which unsolicited advertisements are embodied in automated avatars. A system is provided that includes: a registration system for introducing an advertisement avatar into the virtual universe; a targeting system for targeting a user avatar for delivery of advertising content by the advertisement avatar; a movement system for defining how the advertisement avatar is to move within the virtual universe; and an advertisement delivery system for defining how the advertisement avatar is to deliver the advertising content to the user avatar.

[0005] The drawbacks of the known dialog systems, such as described above, include insufficient speech recognition capabilities for conducting a complex conversation with a user.

[0006] US patent US7203368 discloses a pattern recognition procedure that forms a hierarchical statistical model using HMM (Hidden Markov Model) and CHMM (Coupled Hidden Markov Model). The hierarchical statistical model supports a parent layer having multiple supernodes and a child layer having multiple nodes associated with each supernode of the parent layer. After training, the hierarchical statistical model uses observation vectors extracted from a data set to find a substantially optimal state sequence segmentation. An improvement to this process would be advantageous.

[0007] A more general solution, posing fewer restrictions than solutions based on HMM uses Bayesian networks for speech recognition. Solutions using Bayesian networks, including Dynamic Bayesian Networks (DBN), have been presented in the following publications:

- M. Wester, J. Frankel, and S. King, "Asynchronous articulatory feature recognition using dynamic Bayesian networks" (Proceedings of IEICI Beyond HMM Workshop, 2004),
- J. A. Bilmes and C. Bartels, "Graphical model architectures for speech recognition", IEEE Signal Processing Magazine, vol. 22, pp. 89-100, 2005,
- J. Frankel, M. Wester, and S. King, "Articulatory feature recognition using dynamic Bayesian networks", Computer Speech and Language, vol. 21, no. 4, pp. 620-640, October 2007.
- Stephenson, T. A., Bourlard, H., Bengio, S., & Morris, A. (2000). Automatic speech recognition using dynamic bayesian networks with both acoustic and articulatory variables. In 6th International Conference on Spoken Language Processing: ICSLP 2000 (Interspeech 2000) (No. EPFL-CONF-82579, pp. II-951).

[0008] Speech recognition methods which utilize Bayesian networks are based on modeling of sound duration according to features vector. In DBNs it has become possible to replace a variable representing duration with a variable representing a sound. Nevertheless, all prior art solutions conducted speech analysis in a predefined time range.

[0009] Taking into account the foregoing prior art there is a need to design and implement a speech recognition system and method that would allow improved dialog efficiency between a human and a machine.

DISCLOSURE OF THE INVENTION

[0010] The object of the invention is a computer-implemented method for automatic speech recognition, comprising the steps of registering, by means of an input device, electrical signal representing speech and converting the signal to frequency or time-frequency domain, analyzing the signal in an analysis module based on DBN, configured to generate hypotheses of words (W) and their probabilities on the basis of observed signal features (OA, OV), recognizing a text corresponding to the electrical signal representing speech, on the basis of

certain word (W) hypotheses and their probabilities,. The method is characterized by inputting, to the analysis module, observed signal features which are determined for the signal in frequency or time-frequency domain in at least two parallel signal processing lines for time segments, distinct for each line, and analyzing, in the analysis module, relations between the observed signal features for at least two distinct time segments.

[0011] Preferably, the time segments have a predefined duration.

[0012] Preferably, the time segments depend on the content of speech segments, such as phonemes, syllables, words.

[0013] Preferably, the method further comprises defining, in the analysis module, deterministic and probabilistic relations between variables describing the model, whereas the probabilistic relations are defined at least for linking the observed signal features with a current state (Sti).

[0014] Preferably, the method further comprises analyzing different observed signal features (OA, OV) in a simultaneous way.

[0015] Another object of the invention is a computer-implemented system for speech recognition, comprising an input device for registering an electrical signal representing speech, a module for converting the registered electrical signal representing speech to frequency or time-frequency domain, an analysis module based on a DBN, configured to analyze the signal representing speech and to generate hypotheses of words (W) and their probabilities on the basis of the observed signal features (OA, OV), a module for recognition of text corresponding to the electrical signal representing speech on the basis of the defined hypotheses of words (W) and their probabilities. The system further comprises at least two signal parameterization modules for determining for the analysis module at least two observed signal features in at least two parallel signal processing lines for time segments distinct for each line, wherein the analysis module is configured to analyze dependencies between the observed signal features for at least two distinct time segments.

[0016] The object of the invention is also a computer program comprising program code means for performing all the steps of the computer-implemented method according to the invention when said program is run on a computer, as well as a computer readable medium storing computer-executable instructions performing all the steps of the computer-implemented method according to the invention when executed on a computer.

BRIEF DESCRIPTION OF DRAWINGS

[0017] The object of the invention has been presented in an exemplary embodiment in a drawing, in which:

Fig. 1 presents a block diagram of a system according to the present invention;

Fig. 2 presents a block diagram of the automatic speech recognition process;

Fig. 3 shows modeling of speech with DBNs on parallel time periods of different lengths;

Fig. 4 depicts an example of use of DBN similar to the one shown in Fig. 3 for decoding of sequences of words (a version that has been simplified for exemplary purposes).

MODES FOR CARRYING OUT THE INVENTION

[0018] Fig. 1 presents a block diagram of a system according to the present invention. Such system may be used in interactive advertising or otherwise information providing dialog systems. The dialog shall be as close to a real conversation as possible. Implementation of such assumption is possible due to use of techniques such as pattern recognition, semantic analysis, use of ontology knowledge and natural language generation followed by speech synthesis.

[0019] Dialog systems, in which the present invention may be used, may comprise high quality displays or image projectors. In preferred embodiments the dialog systems may also be equipped with user presence detection or, in more advanced cases, user characteristics detectors such as biometric detectors, face recognition modules and the like. The dialog system may also comprise directional microphones for more efficient acquisition of speech.

[0020] Output information is adjusted to the context of the dialog and determined user preferences. The dialog system preferably also outputs visual avatar or a person' image with which the user talks.

[0021] The dialog system employing speech recognition communicates interactively with a person or a plurality of persons 101. A person 101 inputs questions by speaking to a sound input module, for example to a microphone 102A. The sound registered by the microphone is processed by a speech recognition module 102 and is subsequently delivered to a module for recognizing natural language 103.

[0022] The module for understanding 103 is responsible for interpreting hypotheses of recognitions of statements of a person 101 in a context of anticipated responses in such a manner so that they are understandable for a machine and may be easily and quickly processed. For example, if the system has been implemented at a tourist information spot, the module for understanding 103 based on a list of speech hypotheses with their probabilities has a task of determining whether the speaker seeks a specific place, if he does then what place is it, or a service, information on time, at which public transport operates etc. In a simplest version the module utilizes for this purpose keywords, but here there may be also used a more advanced solution, based on syntax models (e.g. sentence parsers) and/or semantic (e.g. Wordnet or semantic HMM) presented in D. Jurafsky, J.H. Martin, "Speech and Language Processing", Second Edition,

Pearson Education, Prentice Hall, 2009.

[0023] After being processed in the module for understanding of natural language 103, sentences or hypotheses of sentences are passed to a dialog manager module 104 (e.g. such as described in D. Jurafsky, J.H. Martin, "Speech and Language Processing", Second Edition, Pearson Education, Prentice Hall, 2009), which in cooperation with target manager module 106 and targets database 107, by appropriately querying an ontology module 105, determines a response to be presented to a user query.

[0024] The ontology module 105 comprises an ordered knowledge about the universe, for example information on which products are available in certain kinds, what people have bought together with selected one etc. The ontology module may comprise additionally different kinds of data from social services, for example to check whether a friend of the person, with whom the dialog is ongoing, is in the city, which the person visits etc. The ontology module may comprise also any other pragmatic knowledge, systematized in such a manner so that a computer or other machine could process it.

[0025] The target manager module 106 is used to implement, in a computer, known rules of commerce, advertising, negotiations etc., which would direct a specialist person (e.g. commercial employee), whose duties are carried out by the system according to the invention.

[0026] After determining the content of a response, a response in natural language is generated at the module for generating natural language 108 and subsequently in the speech generation module 109. The generated response, in form of speech, is output to the person 101 via a loudspeaker or other output device 109A installed in the system.

[0027] A key element used in the present invention is a computer-implemented module for analysis made of Bayesian networks. Bayesian networks enable modeling of complicated phenomena, in which separate elements may depend on each other. A basic model is created as a directional, acyclic graph, in which nodes represent separate elements of the model (random variables), whereas edges represent dependencies between these elements.

[0028] Additionally, the edges have assigned probability values specifying that one of the events occurs under a condition that another event assumes a particular value. By using Bayes theorem, complex conditional probabilities may be calculated for a particular path of the Bayesian network. These probabilities may be used to infer about values which will be taken by individual elements of the network.

[0029] Each network variable has to be conditionally independent on other variables not connected with it. A graph created in this manner may be interpreted as a compact representation of events, a cumulative probabilities of occurrence of these events as well as assumptions regarding conditional independence between graph's nodes.

[0030] DBNs may be employed for speech recognition. Then, the nodes represent not a single random variable but a sequence of variables. These are interpreted as time series, which allow for speech modeling according to time lapse. Therefore, a plurality of successive observations states give justification to an unambiguous path to a final state.

[0031] The use of standard Bayesian networks is based on anticipating duration of a sound, depend on a vector of articulatory features. The network has a single discrete variable for each feature and a single continuous variable for duration of a sound. The network describes relations between the features. The values of nodes representing features depend on values entered into the network and optionally on other features. A value of a node representing time duration is a hidden layer (as in the HMM), dependent directly only on values received from other nodes.

[0032] Introduction of DBNs allows for replacement of a variable representing duration with a variable representing a sound. The entire network with relations between features is copied in such a way that one of the networks represents a signal analyzed at time t-1 and the next one at time t. Both networks are connected at edges, which have probability values of transition between states that may change in time.

[0033] It is to be noted that the invention is not limited only to a case with two subnetworks. There may be more subnetworks, each subnetwork for a subsequent time moment. Typically, there may be hundreds or thousands of networks. Such a structure may be copied many times to subsequent time moments. Additionally, such local Bayesian network structure may modify itself between distinct times in some cases.

[0034] DBN models may also be used to join information about signals originating from different sources, for example acoustic features and visual features (such as lips movement). Systems of this kind are especially useful in applications for places with difficult acoustic conditions. Low value of Signal to Noise Ratio (SNR) makes that the use of information originating from only acoustic path, in locations such as a street, an airport, a factory etc., significantly decreases the quality of obtained results. Adding information obtained from another signal type, which is not sensitive to the same type of noise, removes the arising difficulties and allows for using speech recognition systems also in such places.

[0035] The inventors have noticed that Bayesian networks pose fewer limitations in comparison to HMM methods when used for speech analysis.

[0036] Fig. 2 presents a block diagram of a speech recognition process. The following description will also reference some features of Fig. 3 that shows modeling of speech with the usage of DBNs on time related periods of different lengths.

[0037] DBNs are used herein for modeling speech in such a way that separate observations represent different time durations - as shown in Fig. 3. These different

time durations may be segments of predefined lengths, e.g. 5ms, 20ms, 60ms, dependent on content of speech segments such as phonemes, syllables, words, or combinations of both types, e.g. 5ms, 20ms, phonemes, words.

[0038] The presented method allows for extraction of different information types and straightforward fusion of acquired features due to use of a DBN model for evaluation of states probability (St1 to St6 in Fig. 3).

[0039] Inferring in DBN is based on two kinds of relations between variables describing a model: deterministic relations (marked in Fig. 3 as straight arrows) and probabilistic relations (marked in Fig. 3 as wave-shaped arrows).

[0040] Deterministic relations are defined on the basis of known facts, e.g. when analyzing a given word Wti there is known the position Wps and the kind of the first phoneme Pti. Then, by knowing that there has occurred or has not occurred a transition Ptr from the phoneme to the next one, there may be determined a position of the current phoneme in a word: Wps at time t + 1 is equal Wps at time t if the transition of phoneme has not been present or is equal to Wps + 1 in case the aforementioned transition has been observed.

[0041] Information regarding transition Wtr from one word to another can be obtained in a similar manner. The occurrence of a transition from the last phoneme of a transcribed word implies a necessity of analysis of another word Wti.

[0042] Another type of relations are probabilistic relations. In order to infer on the basis of variables, between which there exists a probabilistic relation, it is necessary to determine the function defining probability of occurrence of these events (a probability density function - PDF). A relation of this kind is used for linking observed features of a signal with a current state Sti. The preferred PDF functions are Gaussian Mixture Models - GMM.

[0043] Some of the relations may be both deterministic and probabilistic, such as consecutive words Wti. In case a transition from one word to another has not occurred, the relation is deterministic - the word is the same as at the time t-1. In case a transition has occurred, then the next word Wti+1 is determined in a probabilistic manner with a use of knowledge from a language model.

[0044] Inferring in DBN is effected on the basis of observations of acoustic features. However, any observations may be subject to measurement error(s). Introduction of probabilistic relations between related, time-variable observations belonging to the same group (for example OA11, OA23 and OA33 or OV11 and OV23 in Fig. 3) allows to reduce such errors.

[0045] The state Sti and the previous state Sti-1 are used for evaluating probability that the observations are results of speaking a given phoneme (Pt1 to Pt6 in Fig 3).

[0046] The occurrence of a given phoneme also relates probabilistically to a transitory state Ptr. The phoneme Pti, phoneme transition Ptr, position of the phoneme in the word Wps and a transition from the word Wtr allow for evaluating correctness of a hypothesis that the recorded sound contains word W.

[0047] The speech has the characteristic that certain frequency features as well as energy features are almost constant in short time periods. However, in long time periods they vary significantly. Nevertheless, the particular moment when the first and the second situations occur are not defined, hence use of DBN model is very advantageous. Relations between observations in different segments may, but do not have to, be present.

[0048] For example for a variant of a configuration with four time periods they may assume 5ms, 20ms, phonemes and words for parallel analysis. There are possible different model configurations, for example where there are relations between all four ranges but also where there are relations only between the layer of 5ms and 20ms and the layer of phonemes, where there are relations only between the layer of 20ms and the layer of phonemes or where there are relations only between the layer of phonemes and the layer of words.

[0049] Additionally, each of the ranges may have several observation types related to different kinds of features of speech. For example, one of them may be a frequency features vector, another one may be energy and yet another one may be a visual features vector. These may be also acoustic features of the same kind but obtained with different methods (for example WFT (Wavelet-Fourier Transform), MFCC (Mel-Frequency Cepstral Coefficients)) and also acoustic features obtained with the same methods but for different time ranges, for example for a moving window of 20ms, for a moving window of 50ms, both extracted every 10ms.

[0050] Moreover, some ranges may occur only in analysis of a particular kind of features and not be available in other kind (Fig. 3 - observations of acoustic features 1 (308) last for 60ms, observations of acoustic features 2 (310) last for 20ms and observations of visual features 1 (309) last for 30ms).

[0051] There may be more type of features describing the signal used during the analysis, also simultaneously, for example the pitch frequency, formant frequencies, or voiced / unvoiced description of the sound.

[0052] The method presented in Fig. 2 starts at step 201 with an acqusition of a speech signal. The next step 202 is to process the signal to frequency domain by means of e.g. WFT or a time-frequency transform using Short-Time Fourier Transform (STFT). It is possible to apply other transforms allowing for quantitative description of information (like signal energy) comprised in different frequency subbands at different time moments.

[0053] Subsequently, at step 203, the time-frequency spectrum is divided into constant frames, for example 5ms, 20ms, 60ms etc. or segmented according to predefined algorithms, such as presented for example in:

- P. Cardinal, G. Boulianne, and M. Comeau, "Segmentation of recordings based on partial transcriptions", Proceedings of Interspeech, pp. 3345-3348,

2005; or

- K. Demuynck and T. Laureys, "A comparison of different approaches to automatic speech segmentation", Proceedings of the 5th International Conference on Text, Speech and Dialogue, pp. 277-284, 2002; or
- Subramanya, J. Bilmes, and C. P. Chen, "Focused word segmentation for ASR", Proceedings of Interspeech 2005, pp. 393-396, 2005.

[0054] The segmentation module (203) divides the process of spectrum analysis into multiple lines, which will be parameterized independently.

[0055] The number of lines may be different than four as previously described. The example on Fig. 2 employs four separate lines with frames of 5ms - 204a, 20ms - 204b, phoneme - 204c and word - 204d, whereas from each of the lines features are extracted in blocks 204a to 204d, representing speech at a particular time. These parameterization blocks may employ processing algorithms such as MFCC, Perceptual Linear Prediction (PLP), or other such as:

- H. Misra, S. Ikbal, H. Bourlard, and H. Hermansky, "Spectral entropy based feature for robust ASR", Proceedings of ICASSP, pp. I-193-196, 2004; and/or
- L. Deng, J. Wu, J. Droppo, and A. Acero, "Analysis and comparison of two speech feature extraction/compensation algorithms", IEEE Signal Processing Letters, vol. 12, no. 6, pp. 477-480, 2005; and/or
- D. Zhu and K. K. Paliwal, "Product of power spectrum and group delay function for speech recognition", Proceedings of ICASSP, pp. I-125-128, 2004.

[0056] The features obtained from modules 204a to 204d are passed together with observations 201 a, such as a signal energy and a visual features vector, to DBN 205. The DBN model, using its embedded algorithms of approximate inferring for the BN, for example Variational Message Passing, Expectation Propagation and/or Gibbs Sampling, with the use of Dynamic Programming algorithms used in speech recognition, such as Viterbi decoding and/or Baum-Welch, and based on content of the dictionary 206 and the language model 207, for example bigrams of words, determines words hypotheses and calculates their probabilities. In most cases the hypotheses will partially overlap, because the DBN may present different hypotheses for the same time period. The hypotheses may be subsequently processed in a further language model 208 (preferably, more advanced than the first language model used in the DBN), in order to obtain the recognized speech text 209.

[0057] Fig. 3 presents an exemplary DBN structure. Items W 301 denote words, Wtr 302 denote a word transition, Wps 303 denote the position of a phoneme in a particular word, Ptr 304 denote a phoneme transition, Pt 305 denote a phoneme, Spt 306 denote a preceding state, S 307 denote a state, OA1 308 denote observed acoustic features of a first kind in a time window of 60 ms, OV1 309 denote observed visual features of a first kind in a time window of 30 ms, OA2 310 denote observed acoustic features of a second kind in a time window of 20 ms, OA3 311 denote observed acoustic features of a third kind in a time window of 10 ms, while OV2 312 denote observed visual features of a second kind in a time window of 10 ms.

[0058] The arrows represent relations (dependencies) between variables, as previously described. The transitions are defined by conditional probability distributions (CPDs), which are calculated during the training process of Bayesian network, based on training data.

[0059] Fig. 4 depicts an example of use of the DBN shown in Fig. 3, for decoding word sequences. If differs from Fig. 3 in that for speech recognition there has been used one kind of acoustic features of a signal, for two frames of different length. The network presents a process of decoding of the phrase: 'Cat is black' - phonetic transcription: / kæt IZ blæk /. The phoneme state depends on two kinds of observations O1 and 02. The previous state 306 at the time t is an exact copy of state 307 at the time t-1. The analysis is applied to subsequent phonemes of the word 301 depending on the current position in the word 303, occurrences of phoneme transitions to another one 304, the state 306 and the preceding state 307 of the phoneme. The phoneme transition occurs if the value of transition probability is greater than 0.5. The symbols of separate nodes of the Bayesian network from Fig. 3 have been replaced with values of these states. For 302 and 304 they are values: T (True) / F (False) denoting occurrence or lack of occurrence of a transition between subsequent words or subsequent phonemes, respectively. For the position of a phoneme in a word 303 it is an index of the currently analyzed phoneme (1 - 3 for the word 'cat', 1 - 2 for the word 'is', 1 - 4 for the word 'black'). A change of phoneme index occurs only when in the preceding moment of time t-1 a transition of phoneme 304 obtained a value of "T". In addition, the word 301 changes only at the place of occurrence of word transition 302, which is obtained at the moment of phoneme transition 304 from the last index in a particular word. The relation between subsequent words changes in such case from deterministic to probabilistic as a result of using a language model. The exemplary values of bigrams language model (a model utilizing couples of words) are shown in a table above a drawing. Additionally, there have been presented exemplary values of initial word probability in the language model. The technical effect achieved by the simultaneously processing of segments with various time durations and several kinds of features is an increase in speech recognition quality, because one type of phonemes, spoken in various manners, are recognized better at one type of time segments and other requires different type of segments, but determining an appropriate analysis time window for each phoneme kind is complex. Additionally, some features present station-

ary properties, allowing for precise extraction of information at more local time segment, while other require more global time segment. Using the structure as shown in Fig. 3 there may be extracted both kinds of features at once. In traditional systems there are used pieces of information carried only by local features or only by global features. Additionally, for example visual features may have different duration than acoustic ones i.e. for example observation of lips set to speak a sound may last longer or shorter than a particular sound.

[0060] It can be easily recognized, by one skilled in the art, that the aforementioned speech recognition method may be performed and/or controlled by one or more computer programs. Such computer programs are typically executed by utilizing the computing resources in a computing device such as personal computers, personal digital assistants, cellular telephones, receivers and decoders of digital television, information kiosks or the like. Applications are stored in non-volatile memory, for example a flash memory or volatile memory, for example RAM and are executed by a processor. These memories are exemplary recording media for storing computer programs comprising computer-executable instructions performing all the steps of the computer-implemented method according the technical concept presented herein.

[0061] While the invention presented herein has been depicted, described, and has been defined with reference to particular preferred embodiments, such references and examples of implementation in the foregoing specification do not imply any limitation on the invention. It will, however, be evident that various modifications and changes may be made thereto without departing from the broader scope of the technical concept. The presented preferred embodiments are exemplary only, and are not exhaustive of the scope of the technical concept presented herein.

[0062] Accordingly, the scope of protection is not limited to the preferred embodiments described in the specification, but is only limited by the claims that follow.

Claims

1. A computer-implemented method for speech recognition, comprising the steps of:

   - registering (201), by means of an input device (102A), electrical signal representing speech and converting the signal to frequency or time-frequency domain (202),
   - analyzing the signal in an analysis module based on Dynamic Bayesian Network (205), configured to generate hypotheses of words (W) and their probabilities on the basis of observed acoustic (OA) signal features or acoustic (OA) and visual (OV) signal features,
   - recognizing (209), on the basis of certain word (W) hypotheses and their probabilities, a text corresponding to the electrical signal representing speech,

   characterized by:

   - during the recognition process, inputting to the analysis module (205) the observed acoustic (OA) signal features or acoustic (OA) and visual (OV) signal features (308-312) which are determined for the signal in frequency or time-frequency domain (202) in at least two parallel signal processing lines (204a, 204b, 204c, 204d, 201 a), of which at least two signal processing lines are configured to determine acoustic (OA) signal features and at least one line is configured to receive signal segmented in a different manner than the other lines such that at least some frames of at least one line can be shared between different states of the Dynamic Bayesian Network,
   - and analyzing, in the analysis module (205), relations between the observed signal features (308 - 312) for at least two distinct time segments.

2. The method according to claim 1, wherein the time segments have a predefined duration.

3. The method according to claim 1 or 2, wherein the time segments depend on the content of speech segments, such as phonemes, syllables, words.

4. The method according to any of the preceding claims, **characterized by** defining, in the analysis module (205), deterministic and probabilistic relations between variables describing the model, whereas the probabilistic relations are defined at least for linking the observed signal features with a current state (Sti).

5. The method according to any of the preceding claims, **characterized by** analyzing different observed signal features (OA, OV) simultaneously (205).

6. A computer-implemented system for speech recognition, comprising:

   - an input device (102A) for registering an electrical signal representing speech,
   - a module (202) for converting the registered electrical signal representing speech to frequency or time-frequency domain,
   - an analysis module (205) based on a dynamic Bayesian network, configured to analyze the signal representing speech and to generate hypotheses of words (W) and their probabilities on the basis of observed acoustic (OA) signal fea-

tures or acoustic (OA) and visual (OV) signal features,
- a module (209) for recognition of text corresponding to the electrical signal representing speech on the basis of the defined hypotheses of words (W) and their probabilities,

characterized in that the system further comprises:

- at least two signal parameterization modules (204a, 204b, 204c, 204d, 201 a) for determining, during the recognition process, for the analysis module (205) at least two observed acoustic (OA) signal features or acoustic (OA) and visual (OV) signal features (308 - 312) in at least two parallel signal processing lines (204a, 204b, 204c, 204d, 201 a), of which at least two signal processing lines are configured to determine acoustic (OA) signal features and at least one line is configured to receive signal segmented in a different manner than the other lines such that at least some frames of at least one line can be shared between different states of the Dynamic Bayesian Network, ,
- wherein the analysis module (205) is configured to analyze dependencies between the observed signal features (308 - 312) for at least two distinct time segments.

7. A computer program comprising program code means for performing all the steps of the computer-implemented method according to any of claims 1 - 5 when said program is run on a computer.

8. A computer readable medium storing computer-executable instructions performing all the steps of the computer-implemented method according to any of claims 1 - 5 when executed on a computer.

Patentansprüche

1. Computerimplementiertes Verfahren zur Spracherkennung, umfassend die Schritte des:

- Erfassens (201), durch Mittel einer Eingabevorrichtung (102A), von einem die Sprache repräsentierenden elektrischen Signal und des Umwandelns des Signals in einen Frequenz- oder Zeit-Frequenzbereich (202),
- des Analysierens des Signals in einem auf einem dynamischen Bayes-Netzwerk (205) basierendem Analysemodul, das konfiguriert ist, Hypothesen von Wörtern (W) und ihrer Wahrscheinlichkeiten basierend auf beobachteten akustischen (OA) Signalmerkmalen oder akustischen (OA) und visuellen (OV) Signalmerkmalen zu erzeugen,
- des Erkennens (209), basierend auf bestimmten Wort (W)-Hypothesen und ihrer Wahrscheinlichkeiten, eines Textes, der dem die Sprache repräsentierenden elektrischen Signal entspricht,

gekennzeichnet durch:

- Eingeben der beobachteten akustischen (OA) Signalmerkmale oder akustischer (OA) und visueller (OV) Signalmerkmale (308-312) in das Analysemodul (205) während des Erkennungsprozesses, die für das Signal im Frequenz- oder Zeit-Frequenzbereich (202) in mindestens zwei parallelen Signalverarbeitungslinien (204a, 204b, 204c, 204d, 201a) bestimmt werden, von denen mindestens zwei Signalverarbeitungslinien konfiguriert sind, akustische (OA) Signalmerkmale zu bestimmen, und mindestens eine Linie konfiguriert ist, Signalsegmente in einer unterschiedlichen Weise zu empfangen als die anderen Linien, sodass mindestens einige Rahmen der mindestens einen Linie zwischen unterschiedlichen Zuständen des dynamischen Bayes-Netzwerks geteilt werden können,
- und Analysieren der Beziehungen zwischen den beobachteten Signalmerkmalen (308 - 312) im Analysemodul (205) für mindestens zwei verschiedene Zeitsegmente.

2. Verfahren nach Anspruch 1, wobei die Zeitsegmente eine vordefinierte Dauer aufweisen.

3. Verfahren nach Anspruch 1 oder 2, wobei die Zeitsegmente vom Inhalt der Sprachsegmente wie Phoneme, Silben, Wörter abhängen.

4. Verfahren nach einem der vorhergehenden Ansprüche, das **gekennzeichnet ist durch** Definieren im Analysemodul (205) von deterministischen und probabilistischen Beziehungen zwischen Variablen, die das Modell beschreiben, während die probabilistischen Beziehungen mindestens zum Verbinden der beobachteten Signalmerkmale mit einem aktuellen Zustand (Sti) definiert sind.

5. Verfahren nach einem der vorhergehenden Ansprüche, das **gekennzeichnet ist durch** gleichzeitiges (205) Analysieren unterschiedlicher beobachteter Signalmerkmale (OA, OV).

6. Computerimplementiertes System zur Spracherkennung, das Folgendes umfasst:

- eine Eingabevorrichtung (102A) zum Erfassen eines die Sprache repräsentierenden elektrischen Signals,
- ein Modul (202) zum Umwandeln des erfassten

die Sprache repräsentierenden elektrischen Signals in einen Frequenz- oder Zeit-Frequenzbereich,
- ein auf einem Bayes-Netzwerk basierendes Analysemodul (205), das konfiguriert ist, das die Sprache repräsentierende Signal zu analysieren und Hypothesen von Wörtern (W) und ihrer Wahrscheinlichkeiten basierend auf beobachteten akustischen (OA) Signalmerkmalen oder akustischen (OA) und visuellen (OV) Signalmerkmalen zu erzeugen,
- ein Modul (209) zur Erkennung von Text, der dem die Sprache repräsentierenden elektrischen Signal entspricht, basierend auf den definierten Hypothesen von Wörtern (W) und ihrer Wahrscheinlichkeiten,

dadurch gekennzeichnet, dass das System ferner Folgendes umfasst:

- mindestens zwei Signalparametriermodule (204a, 204b, 204c, 204d, 201 a) zum Bestimmen für das Analysemodul (205) während des Erkennungsprozesses von mindestens zwei beobachteten akustischen (OA) Signalmerkmalen oder akustischen (OA) und visuellen (OV) Signalmerkmalen (308-312) in mindestens zwei parallelen Signalverarbeitungslinien (204a, 204b, 204c, 204d, 201a), von denen mindestens zwei Signalverarbeitungslinien konfiguriert sind, akustische (OA) Signalmerkmale zu bestimmen, und mindestens eine Linie konfiguriert ist, Signalsegmente in einer unterschiedlichen Weise zu empfangen als die anderen Linien, sodass mindestens einige Rahmen der mindestens einen Linie zwischen unterschiedlichen Zuständen des dynamischen Bayes-Netzwerks geteilt werden können,,
- wobei das Analysemodul (205) konfiguriert ist, Abhängigkeiten zwischen den beobachteten Signalmerkmalen (308 - 312) für mindestens zwei verschiedene Zeitsegmente zu analysieren.

7. Computerprogramm, das Programmcodemittel zum Ausführen aller Schrittes des computerimplementierten Verfahrens nach einem der Ansprüche 1 - 5 umfasst, wenn das Programm auf einem Computer läuft.

8. Computerlesbares Medium, das computerausführbare Anweisungen umfasst, die alle Schritte des computerimplementierten Verfahrens nach einem der Ansprüche 1 - 5 ausführen, wenn sie auf einem Computer ausgeführt werden.

Revendications

1. Méthode informatisée de reconnaissance de la parole, comprenant les étapes suivantes :

- enregistrement (201) à l'aide d'un dispositif d'entrée (102A) d'un signal électrique représentant la parole, et conversion du signal en domaine fréquentiel ou temps-fréquence (202),
- analyse du signal dans un module d'analyse basé sur un réseau de Bayes dynamique (205), configuré pour générer des hypothèses de mots (W) et leurs probabilités, sur la base de caractéristiques de signaux acoustiques (OA) ou de caractéristiques de signaux acoustiques (OA) et visuels (OV) observées,
- reconnaissance (209), en fonction de certaines hypothèses (W) de mots et leurs probabilités, d'un texte correspondant au signal électrique représentant la parole, **caractérisé par** :
- l'entrée dans le module d'analyse (205), au cours du processus de reconnaissance, des caractéristiques de signaux acoustiques (OA) ou des caractéristiques de signaux acoustiques (OA) et visuels (OV) observées (308-312) qui sont déterminées pour le signal en domaine fréquentiel ou temps-fréquence (202) dans au moins deux lignes de traitement de signal parallèles (204a, 204b, 204c, 204d, 201a), dont au moins deux lignes de traitement de signal sont configurées pour déterminer des caractéristiques de signaux acoustiques (OA) et au moins une ligne est configurée pour recevoir un signal segmenté de façon différente des autres lignes, de sorte qu'au moins certaines trames d'au moins une ligne puissent être partagées entre différents états du Réseau de Bayes dynamique, et
- l'analyse, dans le module d'analyse (205), des relations entre les caractéristiques observées des signaux (308-312) pour au moins deux segments temporels distincts.

2. Méthode selon la revendication 1, les segments temporels ayant une durée prédéfinie.

3. Méthode selon la revendication 1 ou 2, les segments temporels étant fonction du contenu des segments de parole, tels que des phonèmes, des syllabes, des mots.

4. Méthode selon une quelconque des revendications précédentes, **caractérisée par** la définition, dans le module d'analyse (205), de relations déterministes et probabilistes entre des variables décrivant le modèle, tandis que les relations probabilistes sont définies au moins pour la mise en relation du signal observé avec un état actuel (Sti).

5. Méthode selon une quelconque des revendications précédentes, **caractérisée par** l'analyse simultanée (205) de différentes caractéristiques de signaux (OA, OV) observées.

6. Système informatisé de reconnaissance de la parole, comprenant :

- un dispositif d'entrée (102A) pour enregistrer un signal électrique représentant la parole,
- un module (202) de conversion du signal électrique en domaine fréquentiel représentant la parole en fonction de la fréquence ou le temps-fréquence,
- un module d'analyse (205) basé sur un réseau de Bayes dynamique configuré pour analyser le signal représentant la parole et générer des hypothèses de mots (W) et leurs probabilités sur la base de caractéristiques de signaux acoustiques (OA) ou de caractéristiques de signaux acoustiques (OA) et visuels (OV) observées,
- un module (209) pour la reconnaissance de textes correspondant au signal électrique représentant la parole, d'après certaines hypothèses définies de mots (W) et leurs probabilités, **caractérisé en ce que** le système comprend en outre :
- au moins deux modules de paramétrage de signaux (204a, 204b, 204c, 204d, 201a) pour déterminer, au cours du processus de reconnaissance, pour le module d'analyse (205) au moins deux caractéristiques de signaux acoustiques (OA) ou caractéristiques de signaux acoustiques (OA) et visuelles (OV) observées (308 - 312) dans au moins deux lignes de traitement de signaux parallèles (204a, 204b, 204c, 204d, 201a), dont au moins deux lignes de traitement de signaux sont configurées pour déterminer des caractéristiques de signaux acoustiques (OA), et au moins une ligne est configurée pour recevoir un signal segmenté de façon différente des autres lignes, de sorte qu'au moins certaines trames d'au moins une ligne puissent être partagées entre différents états du Réseau de Bayes dynamique,
- le module d'analyse (205) étant configuré pour analyser des dépendances entre les caractéristiques de signaux observées (308 - 312) pour au moins deux segments de temps distincts.

7. Programme informatique comprenant un dispositif de codes de programme pour l'exécution de toutes les étapes de la méthode informatisée selon une quelconque des revendications 1 à 5, lorsque ledit programme est exécuté sur un ordinateur.

8. Support lisible par ordinateur pour stocker des instructions exécutables par ordinateur pour l'exécution de toutes les étapes de la méthode informatisée selon une quelconque des revendications 1 à 5, lorsqu'elles sont exécutées sur un ordinateur.

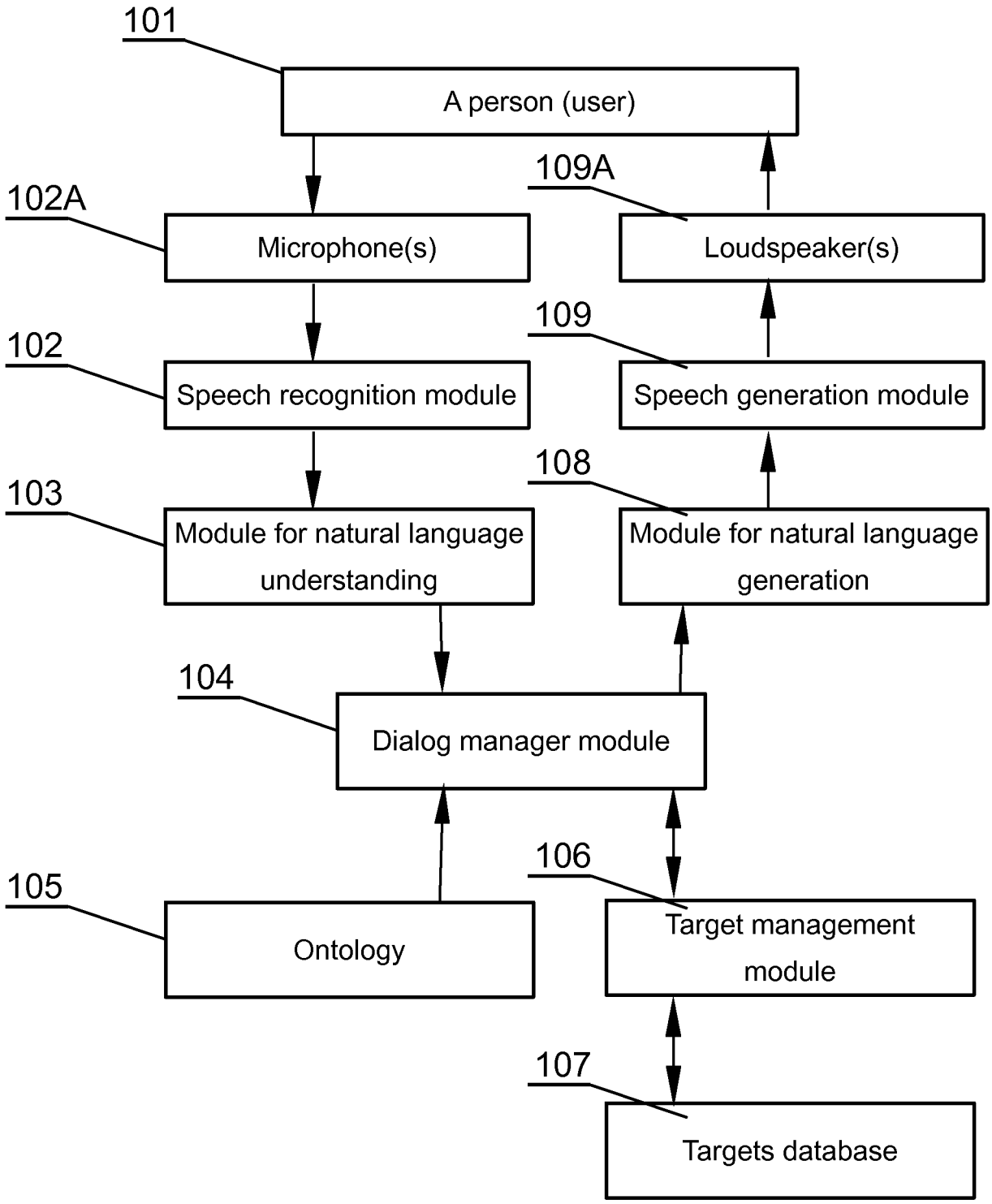
101
A person (user)
102A
Microphone(s)
109A
Loudspeaker(s)
102
Speech recognition module
109
Speech generation module
103
Module for natural language understanding
108
Module for natural language generation
104
Dialog manager module
105
Ontology
106
Target management module
107
Targets database

Fig. 1

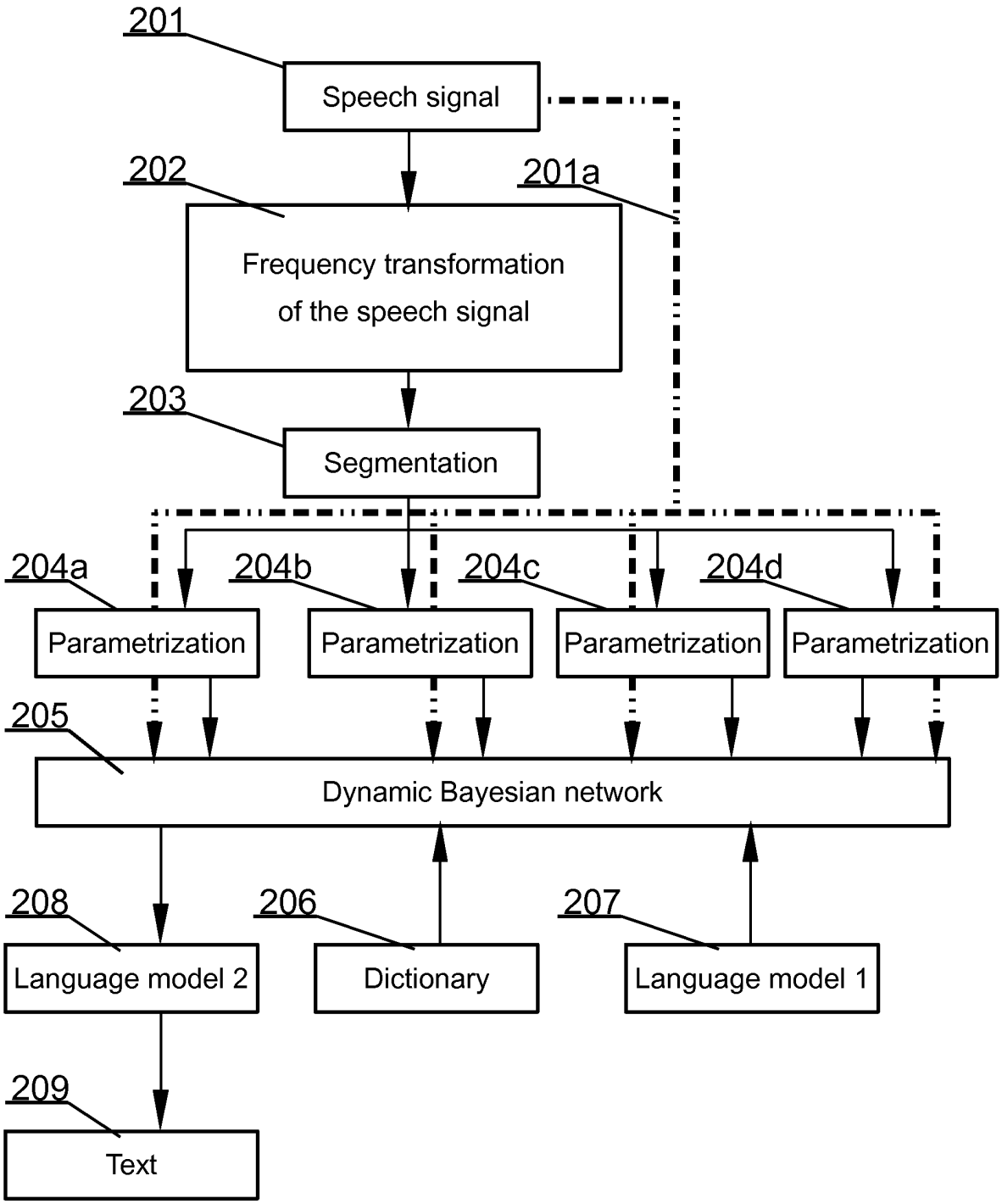
201
Speech signal
202
201a
Frequency transformation of the speech signal
203
Segmentation
204a
204b
204c
204d
Parametrization
Parametrization
Parametrization
Parametrization
205
Dynamic Bayesian network
208
206
207
Language model 2
Dictionary
Language model 1
209
Text

Fig. 2

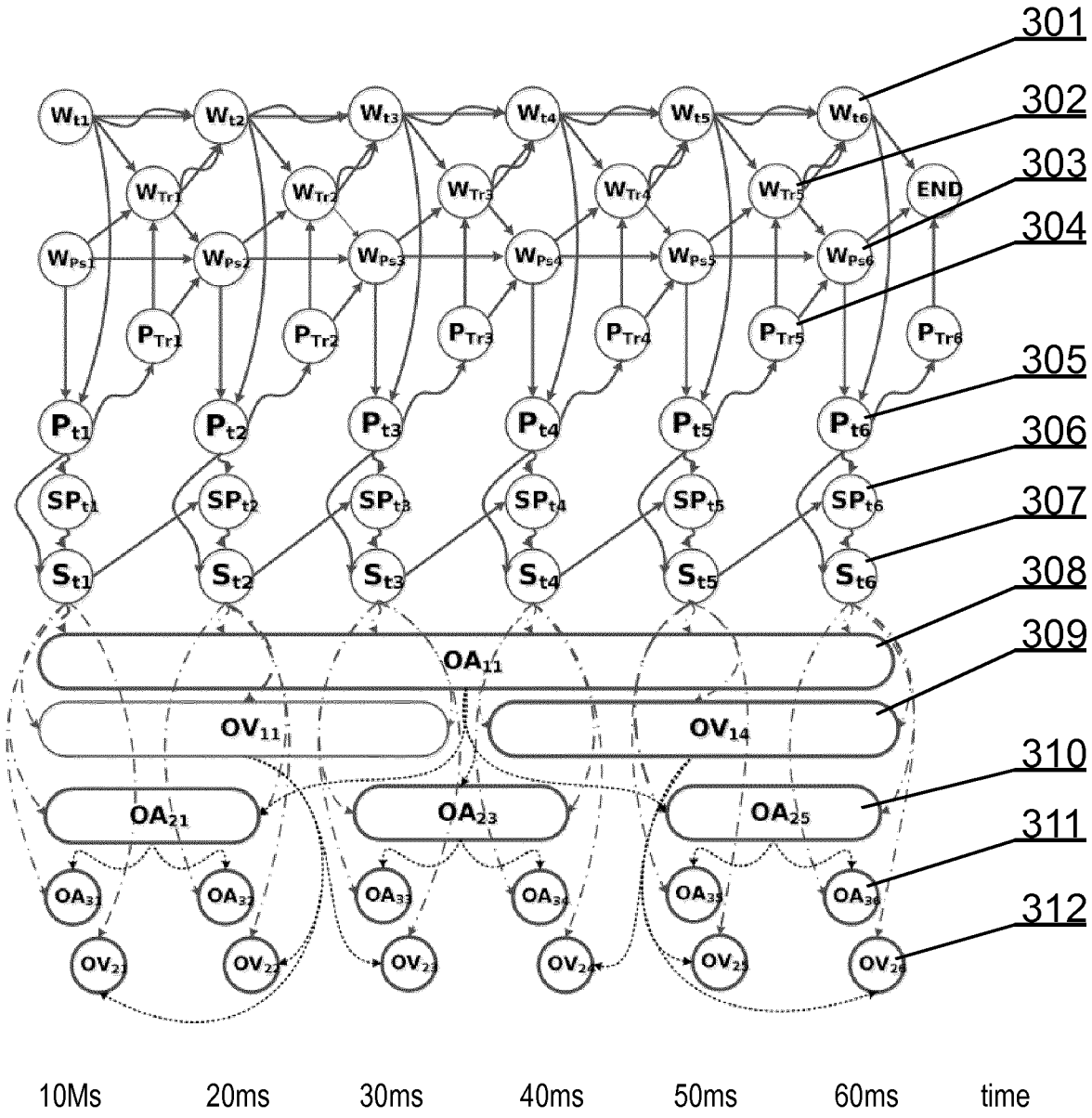
301
302
303
304
305
306
307
308
309
310
311
312
W_{t1} W_{t2} W_{t3} W_{t4} W_{t5} W_{t6}
W_{Tr1} W_{Tr2} W_{Tr3} W_{Tr4} W_{Tr5} END
W_{Ps1} W_{Ps2} W_{Ps3} W_{Ps4} W_{Ps5} W_{Ps6}
P_{Tr1} P_{Tr2} P_{Tr3} P_{Tr4} P_{Tr5} P_{Tr6}
P_{t1} P_{t2} P_{t3} P_{t4} P_{t5} P_{t6}
SP_{t1} SP_{t2} SP_{t3} SP_{t4} SP_{t5} SP_{t6}
S_{t1} S_{t2} S_{t3} S_{t4} S_{t5} S_{t6}
OA_{11}
OV_{11} OV_{14}
OA_{21} OA_{23} OA_{25}
OA_{31} OA_{32} OA_{33} OA_{34} OA_{35} OA_{36}
OV_{21} OV_{22} OV_{23} OV_{24} OV_{25} OV_{26}
10Ms 20ms 30ms 40ms 50ms 60ms time

Fig. 3

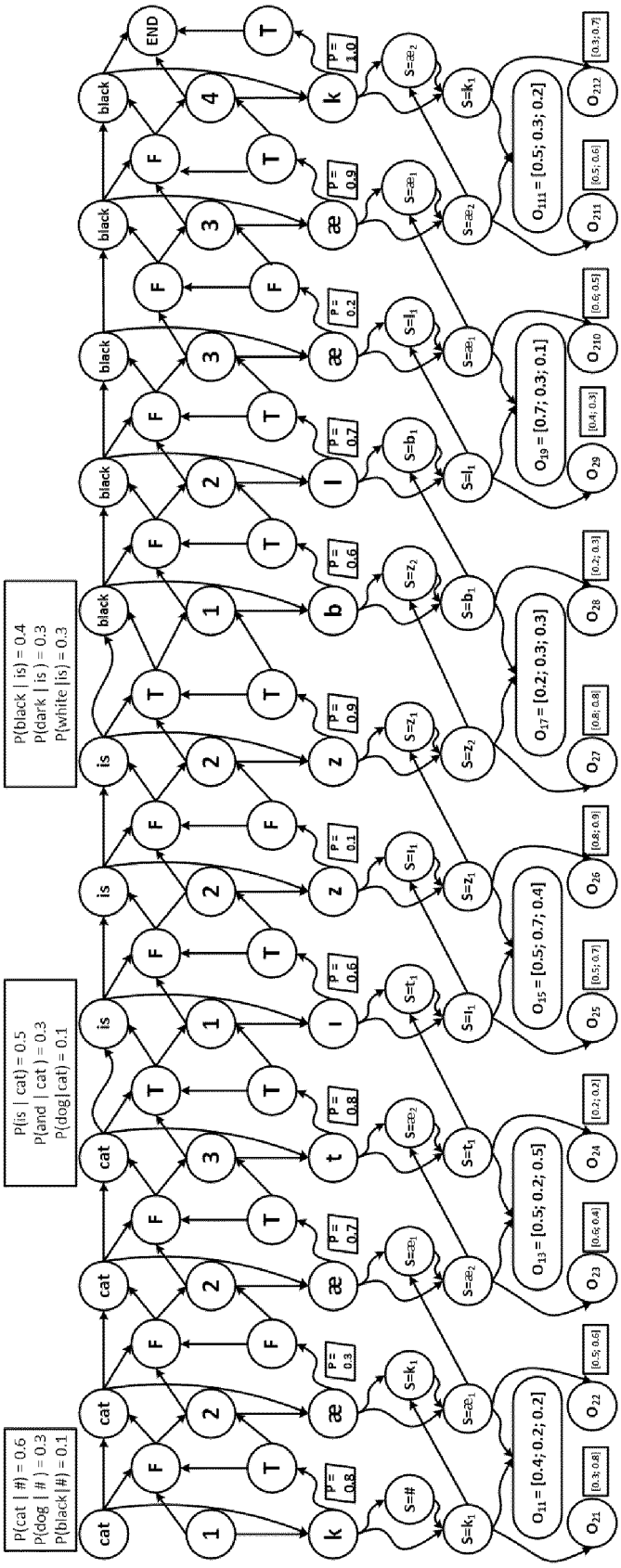
P(cat | #) = 0.6
P(dog | #) = 0.3
P(black | #) = 0.1
P(is | cat) = 0.5
P(and | cat) = 0.3
P(dog|cat) = 0.1
P(black | is) = 0.4
P(dark | is) = 0.3
P(white |is) = 0.3
cat
is
black
END
O11 = [0.4; 0.2; 0.2]
O13 = [0.5; 0.2; 0.5]
O15 = [0.5; 0.7; 0.4]
O17 = [0.2; 0.3; 0.3]
O19 = [0.7; 0.3; 0.1]
O111 = [0.5; 0.3; 0.2]

Fig. 4

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

- US 6256046 B **[0003]**
- US 20080204450 A **[0004]**
- US 7203368 B **[0006]**

Non-patent literature cited in the description

- **M. WESTER ; J. FRANKEL ; S. KING.** Asynchronous articulatory feature recognition using dynamic Bayesian networks. *Proceedings of IEICI Beyond HMM Workshop,* 2004 **[0007]**
- **J. A. BILMES ; C. BARTELS.** Graphical model architectures for speech recognition. *IEEE Signal Processing Magazine,* 2005, vol. 22, 89-100 **[0007]**
- **J. FRANKEL ; M. WESTER ; S. KING.** Articulatory feature recognition using dynamic Bayesian networks. *Computer Speech and Language,* October 2007, vol. 21 (4), 620-640 **[0007]**
- **STEPHENSON, T. A. ; BOURLARD, H. ; BENGIO, S. ; MORRIS, A.** Automatic speech recognition using dynamic bayesian networks with both acoustic and articulatory variables. *In 6th International Conference on Spoken Language Processing: ICSLP 2000,* 2000, II-951 **[0007]**
- Speech and Language Processing. **D. JURAFSKY ; J.H. MARTIN.** Pearson Education. Prentice Hall, 2009 **[0022] [0023]**
- **P. CARDINAL ; G. BOULIANNE ; M. COMEAU.** Segmentation of recordings based on partial transcriptions. *Proceedings of Interspeech,* 2005, 3345-3348 **[0053]**
- **K. DEMUYNCK ; T. LAUREYS.** A comparison of different approaches to automatic speech segmentation. *Proceedings of the 5th International Conference on Text, Speech and Dialogue,* 2002, 277-284 **[0053]**
- **SUBRAMANYA, J. BILMES ; C. P. CHEN.** Focused word segmentation for ASR. *Proceedings of Interspeech 2005,* 2005, 393-396 **[0053]**
- **H. MISRA ; S. IKBAL ; H. BOURLARD ; H. HERMANSKY.** Spectral entropy based feature for robust ASR. *Proceedings of ICASSP,* 2004, I-193-196 **[0055]**
- **L. DENG ; J. WU ; J. DROPPO ; A. ACERO.** Analysis and comparison of two speech feature extraction/compensation algorithms. *IEEE Signal Processing Letters,* 2005, vol. 12 (6), 477-480 **[0055]**
- **D. ZHU ; K. K. PALIWAL.** Product of power spectrum and group delay function for speech recognition. *Proceedings of ICASSP,* 2004, I-125-128 **[0055]**