

(19) World Intellectual Property Organization
International Bureau



(43) International Publication Date
15 November 2007 (15.11.2007)

PCT

(10) International Publication Number
WO 2007/129802 A1

- (51) International Patent Classification:
G10L 15/06 (2006.01) **G10L 15/10** (2006.01)
- (21) International Application Number:
PCT/KR2006/005830
- (22) International Filing Date:
28 December 2006 (28.12.2006)
- (25) Filing Language: Korean
- (26) Publication Language: English
- (30) Priority Data:
10-2006-0042109 10 May 2006 (10.05.2006) KR
- (71) Applicant (for all designated States except US): **KT CORPORATION** [KR/KR]; 206, Jungja-dong, Bundang-gu, Seongnam-city, Gyeonggi-do 463-711 (KR).
- (72) Inventors; and
- (75) Inventors/Applicants (for US only): **RYU, Chang-Sun** [KR/KR]; #14-1103 Cheongsil Apt., Daechi 1-dong, Gangnam-gu, Seoul 135-774 (KR). **KIM, Jae-In** [KR/KR]; 975-3, Sinwol 7-dong, Yangcheon-gu, Seoul 158-846 (KR). **KIM, Hong Kook** [KR/KR]; #623-504 Dongbo

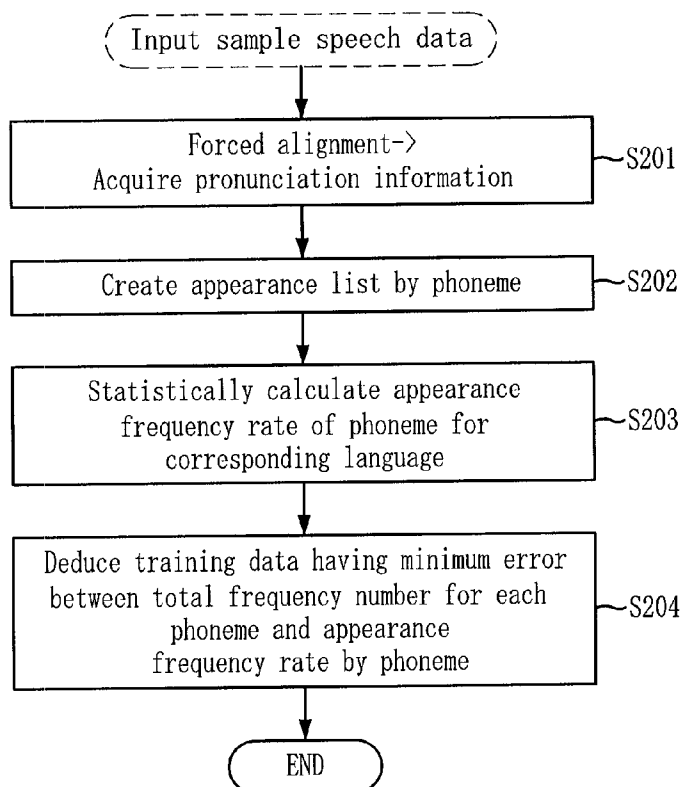
Apt., 968, Yeongtong-dong, Yeongtong-gu, Suwon-si, Gyeonggi-do 443-470 (KR). **OH, Yoo Rhee** [KR/KR]; Gwangju Advanced Institute Of Science, And Technology, Oryong-dong, Buk-gu, Gwangju 500-712 (KR). **LEE, Gil Ho** [KR/KR]; Gwangju Advanced Institute Of Science, And Technology, Oryong-dong, Buk-gu, Gwangju 500-712 (KR).

(74) Agent: **SHINSUNG INTERNATIONAL PATENT & LAW FIRM**; ID Tower #601, Jungdaero 105, (99-7 Garak-dong), Songpa-gu, Seoul 138-805 (KR).

(81) Designated States (unless otherwise indicated, for every kind of national protection available): AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BW, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, 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, KZ, LA, LC, LK, LR, LS, LT, LU, LV, LY, MA, MD, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PG, PH, PL, PT, RO, RS, RU, SC, SD, SE, SG, SK, SL, SM, SV, SY, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

[Continued on next page]

(54) Title: METHOD FOR SELECTING TRAINING DATA BASED ON NON-UNIFORM SAMPLING FOR SPEECH RECOGNITION VECTOR QUANTIZATION



(57) Abstract: Provided is a method for selecting training data based on non-uniform sampling for speech recognition vector quantization. The method includes the steps of : if sample speech data are received, making the speech data subjected to forced alignment to acquire pronunciation information of each phoneme; creating an appearance list by phoneme based on the acquired pronunciation information; statistically calculating an appearance frequency rate of phoneme for a corresponding language depending on the created appearance list by phoneme; and deducing training data having a minimum error between a total frequency number for each phoneme and an appearance frequency rate by phoneme by referring to the calculated statistics.

WO 2007/129802 A1



(84) **Designated States** (*unless otherwise indicated, for every kind of regional protection available*): ARIPO (BW, GH, GM, KE, LS, MW, MZ, NA, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HU, IE, IS, IT, LT, LU, LV, MC, NL, PL, PT, RO, SE, SI, SK, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

Published:

— *with international search report*

DESCRIPTION

METHOD FOR SELECTING TRAINING DATA BASED ON NON-UNIFORM
SAMPLING FOR SPEECH RECOGNITION VECTOR QUANTIZATION

5 TECHNICAL FIELD

The present invention relates to a method for selecting training data for speech recognition vector quantization; and, more particularly, to a method for selecting training data based on non-uniform sampling for
10 speech recognition vector quantization, which selects training data based on non-uniform sampling by considering speech characteristics of language, for example, an appearance frequency by phoneme, in case of selecting training data to be used in training a speech
15 feature vector quantizer.

BACKGROUND ART

Distributed speech recognition (DSR) is a technology which enables a low spec terminal such as a
20 portable telephone to have a speech recognition function. This DSR is constituted by dual processing systems in which the low spec terminal recognizes the feature of a speech signal and a high spec server performs speech recognition based on the feature of the speech signal
25 provided from the low spec terminal.

In general, Mel-Frequency Cepstral Coefficient (MFCC) has been universally used in the speech recognition field. The MFCC represents the form of frequency spectrum expressed in Mel scale as sine wave
30 components, and refers to a speech feature vector (or speech feature parameter) representing speech being received from a user. This MFCC has to be subjected to a quantization process in order to transmit it from the terminal to a server through a communication network. A
35 speech recognition process will be briefly described

below with reference to the DSR system as mentioned above.

At the terminal end, a speech feature vector is first extracted from speech received from a user through MFCC and then quantized to process it in the form
5 suitable for transmitting to a communication network. In other words, MFCCs extracted from speech received from the user, are mapped into vectors having the nearest distance among central vectors of a limited-size codebook. And then, the selected vectors are sent as a bit stream
10 to a server. Therefore, a codebook having a central value for a group with similar values is required for recognition of speech originated from the user, and is used as a reference value for speech feature vector quantization.

15 Meanwhile, the server dequantizes the speech feature vector transmitted from the terminal and recognizes word corresponding to the speech by using HMM (Hidden Markov Model) as a speech model. Here, HMM is a process that models a basic unit, for example, phoneme
20 for speech recognition, wherein words and sentences are created by combining phonemes applied to a speech recognition engine and phonemes registered in DB of the speech recognition engine.

The above-mentioned speech feature vector
25 quantization process is an important factor which determines the performance of the speech recognition system. The design of this speech feature vector quantizer may be defined by the creation of a codebook. Especially, this codebook should be made by extracting
30 training data (training MFCC vector) from numerous speech data and then selecting the most representative values out of the extracted training data as training data for the speech feature vector quantization.

To be more specific, the codebook has central
35 vectors of speech data as indexes and needs a great

amount of training data to enhance the performance of speech recognition. As set forth above, the result of the speech feature vector quantization is a data sequence represented by the codebook index and such a speech
5 feature vector quantization gives the nearest codebook index by distance calculation between vectors.

However, most of speech recognition technologies to date propose only a processing algorithm of the speech feature vector quantizer, but do not propose an algorithm
10 for selecting training data actually required for training the speech feature vector quantizer.

For example, in the speech feature vector quantizer, it is current situation that the prior art relating to the selection of the training data uses only
15 arbitrary unplanned sample training data selected by an administrator in advance.

Further, as described above, the recognition performance of HMM is excellent and thus most of speech recognition systems adopt this HMM. However, in case of
20 such HMM, if the reliability of training data is low, the server has a difficulty in estimating a precise acoustic model and also needs an enormous amount of prior knowledge about phoneme context. For example, if the reliability of training data is high in the quantization
25 of speech feature vector at the terminal end, the server does not require a separate speech recognition process but performs only a dequantization of the speech feature vector through the existing technology.

Therefore, there is an urgent need for a
30 technology capable of quantizing speech feature vector with a smaller number of bits. It will easily be understood by those skilled in the art that the development of an algorithm of selecting training data used as training data in the speech feature vector
35 quantization to do so should be preceded above all things.

DISCLOSURE

TECHNICAL PROBLEM

An embodiment of the present invention is directed
5 to provide a method for selecting training data based on
non-uniform sampling for speech recognition vector
quantization, which selects training data based on non-
uniform sampling by considering speech characteristics of
language, for example, an appearance frequency by phoneme,
10 in case of selecting training data to be used in training
a speech feature vector quantizer.

Other objects and advantages of the present
invention can be understood by the following description,
and become apparent with reference to the embodiments of
15 the present invention. Also, it is obvious to those
skilled in the art of the present invention that the
objects and advantages of the present invention can be
realized by the means as claimed and combinations thereof.

20 TECHNICAL SOLUTION

In accordance with an aspect of the present
invention, there is provided a method for selecting
training data based on non-uniform sampling for speech
recognition vector quantization, the method including the
25 steps of: if sample speech data are received, making the
speech data subjected to forced alignment to acquire
pronunciation information of each phoneme; creating an
appearance list by phoneme based on the acquired
pronunciation information; statistically calculating an
30 appearance frequency rate of phoneme for a corresponding
language depending on the created appearance list by
phoneme; and deducing training data having a minimum
error between a total frequency number for each phoneme
and an appearance frequency rate by phoneme by referring
35 to the calculated statistics.

ADVANTAGEOUS EFFECTS

As mentioned above and will be described below, the present invention can more improve the speech
5 recognition performance under the same condition of bit number necessary for the speech feature vector quantization, compared with the existing method using the unplanned sample of training data.

In addition, the present invention exhibits the
10 same speech recognition performance compared with the existing system which does not perform the speech feature vector quantization process, and prevents the lowering of the speech recognition performance that may be caused by going through the speech feature vector quantization
15 process.

Further, the present invention can enhance the reliability of training data in speech recognition and thus can guarantee the superior speech recognition accuracy by estimating a precise acoustic model through
20 HMM, even through an existing speech recognition device which is not provided with a separate processor or equipment.

BRIEF DESCRIPTION OF THE DRAWINGS

25 Fig. 1 is a block diagram illustrating a structure of one example of a DSR system to which the present invention is applied.

Fig. 2 is a flowchart describing a training data selection method based on non-uniform sampling in a
30 speech feature vector quantization in accordance with a preferred embodiment of the present invention.

Fig. 3 is an explanatory diagram of one example of a speech recognition process used in the present invention.

35 Fig. 4 is a flowchart describing one example of a

training process of creating an acoustic model in Fig. 3.

Fig. 5 is a graph representing a speech recognition performance of a speech feature vector quantizer in which the present invention is applied to
5 Korean.

Fig. 6 is a graph representing a speech recognition performance of a speech feature vector quantizer in which the present invention is applied to English.

10

BEST MODE FOR THE INVENTION

The advantages, features and aspects of the invention will become apparent from the following description of the embodiments with reference to the
15 accompanying drawings, which is set forth hereinafter, and thus the invention will easily be carried out by those skilled in the art to which the invention pertains. Further, in the following description, well-known arts will not be described in detail if it seems that they
20 could obscure the invention in unnecessary detail. Hereinafter, a preferred embodiment of the present invention will be described in detail with reference to the accompanying drawings.

Fig. 1 is a block diagram illustrating a structure
25 of one example of a DSR system to which the present invention is applied.

As shown in Fig. 1, the DSR system to which the present invention is applied includes a low spec terminal (client) such as a portable telephone and a high spec
30 server.

The terminal includes an A/D converter 11, a speech feature vector extractor 12, a speech feature vector quantizer 13 and so on. The server includes a speech feature vector dequantizer 21, a speech recognizer
35 22 and so on. Other basic components are omitted here

for simplicity. In addition, in the following description, although the DSR system is described for illustration, it will easily be understood by those skilled in the art that the present invention can also be applied to a single speech recognition system, for example, other speech recognition device comprised of an A/D converter, a speech feature vector extractor, a speech feature vector quantizer/dequantizer, and a speech recognizer.

10 The present invention is characterized by selecting training data based on non-uniform sampling by considering speech characteristics of languages (e.g., Korean, English, etc.), for example, an appearance frequency by phoneme, in case of selecting training data to be used in training (or designing) the speech feature vector quantizer 13.

That is, the present invention employs statistical characteristics of language in selecting training data to quantize a speech feature vector represented by MFCC. For example, speech characteristics are represented differently depending on languages and training data are selected based on the speech characteristics as mentioned above. On the basis of the training data, a codebook is created and then applied to the speech feature vector quantizer 13 so as to enhance the accuracy of speech recognition.

The components of the DSR system will be roughly described below, prior to explaining the training data selection algorithm of the present invention.

30 The A/D converter 11 converts a speech signal of analog form received from a user into speech data of digital form.

The speech feature vector extractor 12 extracts a speech feature vector representing speech characteristics from the speech data converted by the A/D converter 11.

The speech feature vector quantizer 13 quantizes the speech feature vector extracted from the speech feature vector extractor 12 in the form of a bit stream suitable for the communication network based on the training data previously trained through the training data. In addition, a quantization process applied to the speech feature vector quantizer 13 uses a split vector quantization scheme proposed by the standard "ETSI ES 201 108" of ETSI (European Telecommunications Standards Institute).

The speech feature vector dequantizer 21 dequantizes the bit stream data transmitted from the terminal end.

The speech recognizer 22 recognizes the data dequantized by the speech feature vector dequantizer 21, that is, the speech feature vector based on HMM to output speech recognition results such as a word, a sentence, etc. In addition, the server can also perform the function of reconstructing speech by using the speech feature vector and reproducing it.

In this DSR system, the speech feature vector is extracted through MFCC and represented as 14th order coefficients, which are subjected to a quantization process. That is, it is first noted that the MFCC vector is not represented as a particularly fixed order of coefficients, but represented as 14th or 15th order coefficients since the speech recognition performance is excellent in case of expressing it as such order of coefficients.

Hereinafter, a training data selection method based on non-uniform sampling proposed by the present invention will be described in detail with reference to Fig. 2. Further, a speech recognition process and an acoustic model generation process used in the present invention will be described with reference to Figs. 3 and

4, and a speech recognition performance when the present invention is applied to Korean and English will be described with reference to Figs. 5 and 6.

Fig. 2 is a flowchart illustrating the training
5 data selection method based on non-uniform sampling for
speech feature vector quantization in accordance with a
preferred embodiment of the present invention.

First of all, when sample speech data used in training the speech feature vector quantizer is received, the speech data are subjected to forced alignment to acquire pronunciation information of each phoneme at step S201. That is, an utterance sentence, a phonetic transcription, a phoneme type, and a start time and an end time by phoneme are obtained by making the speech data subjected to the forced alignment. The following Table 1 represents the result which is obtained by having the sample speech data subjected to the forced alignment.

Table 1

[illegible]

Phoneme type	n(ㄴ)	a(ㅏ)	o(ㅓ)	g(ㄱ)	o(ㅓ)				
Phoneme start time(s)	1.990	2.070	2.160	2.230	2.280				
Phoneme end time(s)	2.070	2.160	2.230	2.280	2.310				
Phoneme type	i(ㅣ)	d(ㄷ)	z(ㅈ)	o(ㅓ)	sp	sil			
Phoneme start time(s)	2.310	2.370	2.420	2.610	2.800	2.860			
Phoneme end time(s)	2.370	2.420	2.610	2.800	2.860	3.050			

In Table 1, it can be seen that the sample speech data represent speech which pronounces “그리고 계속해서 사람들이 뛰어 나오고 있죠 (and people run out continuously)”

and transcribes as
“ㄱ-ㅡ | ㅈㅊㅅㅣ ㅣㅏㅑㅣㅣㅓㅣㅔㅣㅕㅣㅖㅣㅗㅛㅜㅝ | ㅠㅒㅣㅣㅌㅣㅍㅇㅣㅣㅘㅙ”,
show a start time and an end time by phoneme.

Thereafter, an appearance list by phoneme is created on the basis of the pronunciation information acquired by the forced alignment process at step S202. The following Table 2 represents an appearance list by phoneme in the result of executing the forced alignment for speech data stored in a Korean DB used in verifying the present invention. The last two rows represent a frequency by phoneme and a total phoneme number of training data.

Table 2

			Phoneme	
			g()	i()
List by phoneme	1	file name	"speech_data_dir /speech_file1"	"speech_data_dir /speech_file1"
		start time	0.060	0.260
		end time	0.013	0.290
	2	file name	"speech_data_dir /speech_file1"	"speech_data_dir /speech_file1"
		start time	0.290	1.630
		end time	0.320	1.730
	3	file name	"speech_data_dir /speech_file1"	"speech_data_dir /speech_file1"
		start	0.490	2.310

		time end time	0.610		2.370
	4	file name	"speech_data_dir /speech_file1"		"speech_data_dir /speech_file2"
		start time	2.230		0.074
		end time	3.380		0.082
	...	file name			
		start time	...		...
		end time			
Frequency	by	total	150768	...	155210
phoneme					
Total		phoneme	2086617		
appearance		number			

Next, an appearance frequency rate for phoneme according to the corresponding language is statistically calculated based on the created appearance list by phoneme at step S203. Here, an appearance frequency rate for a certain pronunciation among pronunciation sequences is analyzed on the basis of the pronunciation dictionary, in statistically calculating the appearance frequency rate for phoneme. For example, in order to analyze the phoneme appearance frequency of Korean, a phoneme appearance frequency for each consonant and vowel is analyzed based on the pronunciation dictionary data such as "a study on an appearance frequency of voice appearing in a usual conversation of adult", and "a quantitative linguistic study on function load of Korean phonemes and a phoneme chain". On the other hand, a phoneme appearance frequency analysis for each consonant and vowel is performed depending on the pronunciation dictionary data such as CMU dictionary and the like in order to analyze the phoneme appearance frequency of Korean.

The following Table 3 represents an appearance frequency rate of Korean phoneme, and the following Table 4 represents an appearance frequency rate of English

phoneme.

Table 3

Phoneme	ㅅ	ㅈ	ㅇ	ㅊ	ㅋ	ㅌ	ㄴ	ㅍ	ㅊ	ㅌ
Frequency(%)	1.11	10.24	2.18	0.09	0.12	4.68	0.01	1.73	0.66	2.61
Phoneme	ㄱ	ㅊ	ㅋ	ㅌ	ㅍ	ㅊ	ㅌ	ㅍ	ㅊ	ㅌ
Frequency(%)	9.23	0.71	2.28	0.09	5.03	0.24	3.06	0.93	0.23	1.12
Phoneme	ㅊ	ㅌ	ㅍ	ㅊ	ㅌ	ㅍ	ㅊ	ㅌ	ㅍ	ㅊ
Frequency(%)	1.53	3.33	5.80	0.01	2.88	0.45	0.16	0.58	0.46	7.28
Phoneme	ㅌ	ㅍ	ㅊ	ㅌ	ㅍ	ㅊ	ㅌ	ㅍ	ㅊ	ㅌ
Frequency(%)	0.24	3.95	0.03	6.98	0.44	1.69	3.99	0.53	2.16	11.22

Table 4

Phoneme	S	UW	T	N	K	Y	Z	AO	AY	SH
Frequency(%)	5.90	1.13	5.67	7.12	4.98	0.59	3.20	1.31	1.32	1.02
Phoneme	W	NG	EY	B	CH	OY	JH	D	ZH	G
Frequency(%)	1.04	1.15	1.56	2.45	0.58	0.14	0.74	3.80	0.06	1.59
Phoneme	UH	F	V	ER	AA	M	IH	DH	L	AH
Frequency(%)	0.26	1.60	1.26	3.42	2.88	3.44	5.89	0.06	5.80	8.33
Phoneme	P	OW	AW	HH	AE	R	TH	IY	EH	
Frequency(%)	2.30	2.20	0.39	1.09	2.54	5.40	0.34	4.05	3.21	

5 In Table 2 and Table 3 (or Table 4 in English),
training data having a minimum error between the total
frequency (Table 2) for each phoneme and the appearance
frequency rate by phoneme (Table 3) are derived at step
S204. An example of this training data deriving process
10 (step S204) will be described later.

Through these steps S201 to S204, the selected
training data are determined as the training data of the
speech feature vector quantizer 13. The training data
thus determined are used as parameter for creation of a
15 codebook. Here, any algorithms such as K-means algorithm
can be used as the algorithm used in creating the
codebook.

In order to help understand the training data
selection algorithm of the present invention mentioned
20 above, for example, if an English frequency rate of a
phoneme "/a/" in the pronunciation dictionary is 2%, the
number of times of generation of the phoneme "/a/" in the
sample training data (sample speech data) is "40" and the
total number of phonemes generated in the sample training
25 data is "1,000", the frequency rate at which the phoneme

"/a/" is to be generated in the sample training data is 4%. Then, it is subjected to the forced alignment and then only "1, 3, 5,..., 35, 37, and 39th data" are selected in the created sample training data list for the phoneme
5 "/a/" to use (derive) them as the training data of the speech feature vector quantizer.

For example, in order to minimize a difference between an appearance frequency rate for a specific phoneme in the pronunciation dictionary and an appearance
10 frequency rate of the corresponding phoneme with respect to a total phoneme number in the sample training data (that is, 4%→2%, 1/2 difference), 1/2 sample training data ("1, 3, 5,..., 35, 37, and 39th data") out of the total 40 sample training data are selected as the
15 training data for training the speech feature vector quantizer.

Fig. 3 is an explanatory diagram of one example of the speech recognition process used in the present invention, and Fig. 4 is a flowchart describing one
20 example of the training process of creating the acoustic model in Fig. 3.

Fig. 3 represents the speech recognition process using HMM, which carries out the processes of extracting a speech feature 301 from user's speech, and searches an
25 acoustic model 303, a language model 304 and a pronunciation dictionary 305 for the extracted speech feature, to recognize a word and a sentence corresponding to the speech through a pattern matching 302.

The speech feature extraction 301 employs a scheme
30 proposed by the standard "ETSI ES 201 108" of ETSI. In other words, the speech feature is extracted from the speech data through MFCC and the speech feature vector is formed as 14th order coefficients. For this speech feature vector, a word sequence with a maximum
35 probability value is searched by the pattern matching

that uses the acoustic model 303, the language model 304 and the pronunciation dictionary 305.

The acoustic model 303 uses HMM, and particularly, the present invention uses the phoneme model according to language as the acoustic model. A training process for creation of this phoneme model will be described below with reference to Fig. 4.

First, a monophone-based model which is a phoneme-independent model is created by using the speech feature vector extracted from the training data selected according to the present invention with reference to Fig. 2 at step S401.

Subsequently, it is subjected to forced alignment on the basis of the created monophone-based model, to newly create a phoneme label file at step S402.

In the meantime, a triphone-based model which is a phoneme-dependent model is created by expanding the created monophone-based model at step S403.

Next, a state-tying is made by considering the fact that the amount of training data for the created triphone-based model is small at step S404.

Then, the model resulting from the state-tying is subjected to mixture density increase, to lastly create an acoustic model at step S405.

Meanwhile, the language model 304 shown in Fig. 3 uses a statistical-based scheme. Here, the statistical-based scheme refers to a scheme which statistically estimates a probability value of a possible word sequence from the database (DB) of speech originated under a given circumstance. A representative model of language models of the statistical-based method is an n-gram. This n-gram method approximates the probability of word sequence as multiplication of n-number of its previous conditional probabilities for its use, and uses bigram in Fig. 3.

The pronunciation dictionary 305 uses the

pronunciation dictionary provided by "CleanSent01" of SiTEC (Speech Information Technology & Industry Promotion Center) in case of Korean, and "CMU dictionary V.0.6" provided by Carneige Mellon University in case of English.

5 On the other hand, the speech DB in Fig. 3 uses "read sentence speech DB (CleanSent01)" provided by SiTEC in case of Korean, and uses "AURORA 4 DB(Wall Street Journal)" established by ETSI in case of English.

10 Fig. 5 is a graph representing a speech recognition performance of a speech feature vector quantizer in which the present invention is applied to Korean, and Fig. 6 is a graph representing a speech recognition performance of a speech feature vector quantizer in which the present invention is applied to
15 English.

 In Fig. 5, the left graph represents an average word error rate for Korean in case of training a speech feature vector quantizer using unplanned sample training data, and the right graph represents an average word
20 error rate for Korean in case of training a speech feature vector quantizer using training data selected according to an appearance frequency by phoneme presented by the present invention.

 In Fig. 6, the left graph represents an average
25 word error rate for English in case of training a speech feature vector quantizer using unplanned sample training data, and the right graph represents an average word error rate for English in case of training a speech feature vector quantizer using training data selected
30 according to an appearance frequency by phoneme proposed by the present invention.

 The following is a description about the speech recognition performance experiment shown in each of Figs. 5 and 6.

35 In English, multi-condition training data of "Wall

Street Journal DB (WSJO)" is used as sample training data, and 7,138 speech data which are recorded by "Sennheiser short distance microphone" and several types of long distance microphone are used as speech input condition.

5 These speech data are stored at a sampling rate of 16 kHz.

Further, among evaluation data of "WSJO" corresponding to 5,000-word class, "Set1 to Set7" which are part of 14 data sets for multi-condition evaluation are used for speech recognition performance evaluation.

10 Here, each set is composed of 330 speech utterances, which are recorded from 8 speakers and issued from about 40 utterances per speaker.

Also, 39 order speech feature vectors are used. For this, 12 order MFCC and log energy are extracted and
15 their difference and accelerated coefficients are also used. Here, the DSR front-end standard algorithm proposed by ETSI is used in the extraction of the speech feature vector.

Further, cepstral mean normalization and energy
20 normalization are employed for the speech feature vector used in the training.

Meanwhile, the acoustic model employs "left-to-right model" having 3 states, and the context-independent 4-mixture density and "cross-word triphone model". This
25 acoustic model uses "HTK version 3.2 toolkit" for training. For example, after extending to a triphone-based acoustic model starting from 41 monotone based acoustic models where two pause models are contained, a state-tying using a decision tree is conducted, to
30 thereby reduce the number of states. The total number of acoustic models and the total number of states are 8,360 and 5,356, respectively.

After setting up the speech recognition system in the above condition, the speech recognition performance
35 in case of performing the speech feature vector

quantization process, that in case of training the speech feature vector quantizer using the unplanned sample training data, and that in case of training the speech feature vector quantizer using the training data selected according to the appearance frequency by phoneme proposed by the present invention are compared as follows.

The following Table 5 represents a word error rate in the speech recognition performance without the speech feature vector quantization process. In Table 5, it can be seen that an average word error rate for 7 sets is 21.8 %.

Table 5

Set 1	Set 2	Set 3	Set 4	Set 5	Set 6	Set 7	Avg
14.3	17.6	22.1	27.0	25.1	23.1	23.5	21.8

Thereafter, a word error rate for a method in which a codebook is created from the unplanned training data and applied to the speech feature vector quantizer and that for a method in which a codebook is created from the training data selected by using the algorithm proposed by the present invention and applied to the speech feature vector quantizer are compared as follows.

The following Table 6 represents a word error rate for a method in which a codebook is created from the unplanned training data and applied to the speech feature vector quantizer, and Table 7 represents a word error rate for a method in which a codebook is created from the training data selected by using the algorithm suggested by the present invention and applied to the speech feature vector quantizer.

Table 6

Set 1	Set 2	Set 3	Set 4	Set 5	Set 6	Set 7	Avg
17.8	18.9	25.1	29.4	26.5	24.8	25.9	24.1

Table 7

Set 1	Set 2	Set 3	Set 4	Set 5	Set 6	Set 7	Avg
13.6	16.8	23.0	27.6	24.9	21.8	24.5	21.7

In Tables 6 and 7, it can be seen that the existing method using the unplanned training data shows that the average word error rate for 7 sets is 24.1%, while the
5 training data selection method according to the phoneme appearance frequency provided by the present invention shows that the average word error rate for 7 sets is 21.7%.

In other words, it can be confirmed that the word
10 error rate resulting from the algorithm presented by the present invention is decreased by about 10%, compared with that for the existing unplanned training data selection algorithm. This is almost identical to 21.8% which is the average word error rate of the method
15 without the speech feature vector quantization process. It will easily be understood by those skilled in the art that the present invention does not lower the speech recognition performance by way of going through the speech feature vector quantization process.

20 As described above, the training data selection method provided by the present invention can also be applied to any devices or systems having other speech feature vector quantizers as well as the DSR system shown in Fig. 1. For example, the training data selection
25 method of the present invention can also be applied to a quantizer of linear prediction coefficient (LPC) used in speech encoding. It will easily be appreciated by those skilled in the art that the training data selection method of the present invention can be applied to
30 multimedia processing systems of image, audio, etc. (commonly called "speech processing device"), in view of characteristics of vector to be quantized.

The method of the present invention as mentioned above may be implemented by a software program that is
35 stored in a computer-readable storage medium such as CD-

ROM, RAM, ROM, floppy disk, hard disk, optical magnetic disk, or the like. This process may be readily carried out by those skilled in the art; and therefore, details of thereof are omitted here.

5 While the present invention has been described with respect to the particular embodiments, it will be apparent to those skilled in the art that various changes and modifications may be made without departing from the spirit and scope of the invention as defined in the
10 following claims.

WHAT IS CLAIMED IS:

1. A method for selecting training data based on non-uniform sampling for speech recognition vector
5 quantization, the method comprising the steps of:
if sample speech data is received, making the speech data subjected to forced alignment to acquire pronunciation information of each phoneme;
creating an appearance list by phoneme based on the
10 acquired pronunciation information;
statistically calculating an appearance frequency rate of phoneme for a corresponding language depending on the created appearance list by phoneme; and
deducing training data having a minimum error
15 between a total frequency number for each phoneme and an appearance frequency rate by phoneme by referring to the calculated statistics.
2. The method of claim 1, wherein the
20 pronunciation information contains an utterance sentence, a phonetic transcription, a phoneme type, and a start time and an end time by phoneme.
3. The method of claim 1, wherein the appearance
25 list by phoneme contains a frequency number by phoneme and a total phoneme number of training data.
4. The method of claim 1, wherein the speech
feature vector is represented by MFCC (Mel-Frequency
30 Cepstral Coefficient).
5. The method of claim 1, wherein the language is Korean.

6. The method of claim 1, wherein the language is English.

7. The method of claim 1, wherein the training
5 data selected from the training data deduction is used as training data for creating an acoustic model in HMM (Hidden Markov Model).

8. The method of claim 1, wherein a rate for a
10 target pronunciation out of pronunciation sequences is analyzed based on a pronunciation dictionary of the corresponding language in the phoneme appearance frequency rate calculation.

9. The method of claim 1, wherein, in the
15 training data deduction, the training data are selected from sample training data as many as a number corresponding to a difference between an appearance frequency rate for a specific phoneme in the
20 pronunciation dictionary and an appearance frequency rate of the corresponding phoneme with respect to a total phonemic number in the sample speech data.

10. A computer-readable recording medium for
25 storing a program implementing a method for selecting training data in a speech processing apparatus with a processor, the method comprising the steps of:

if sample speech data is received, making the
speech data subjected to forced alignment to acquire
30 pronunciation information of each phoneme;

creating an appearance list by phoneme based on the
acquired pronunciation information;

statistically calculating an appearance frequency
rate of phoneme for a corresponding language depending on
35 the created appearance list by phoneme; and

deducing training data having a minimum error between a total frequency number for each phoneme and an appearance frequency rate by phoneme by referring to the calculated statistics.

1/4

FIG. 1

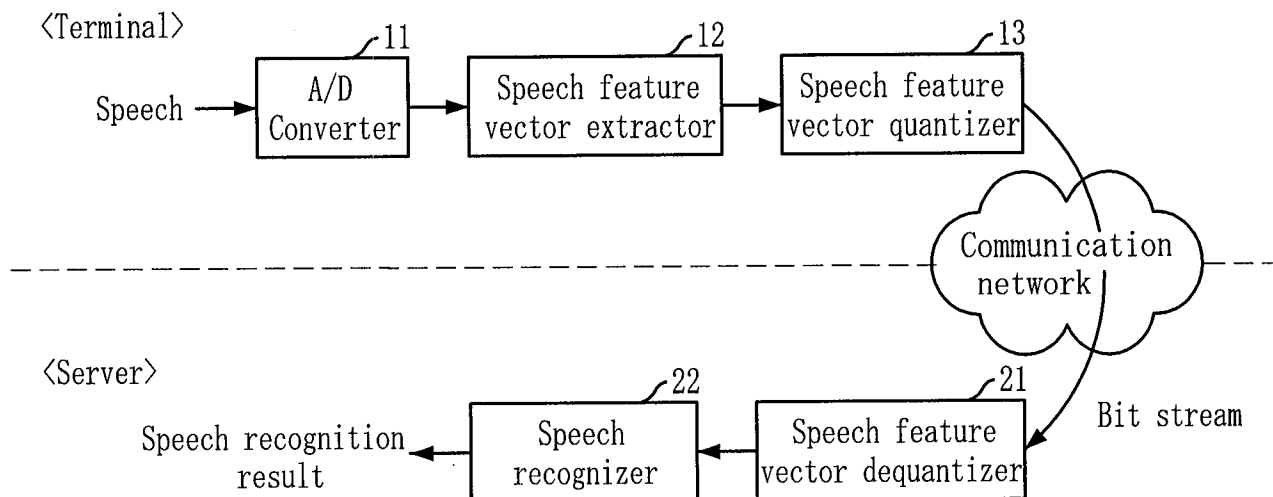
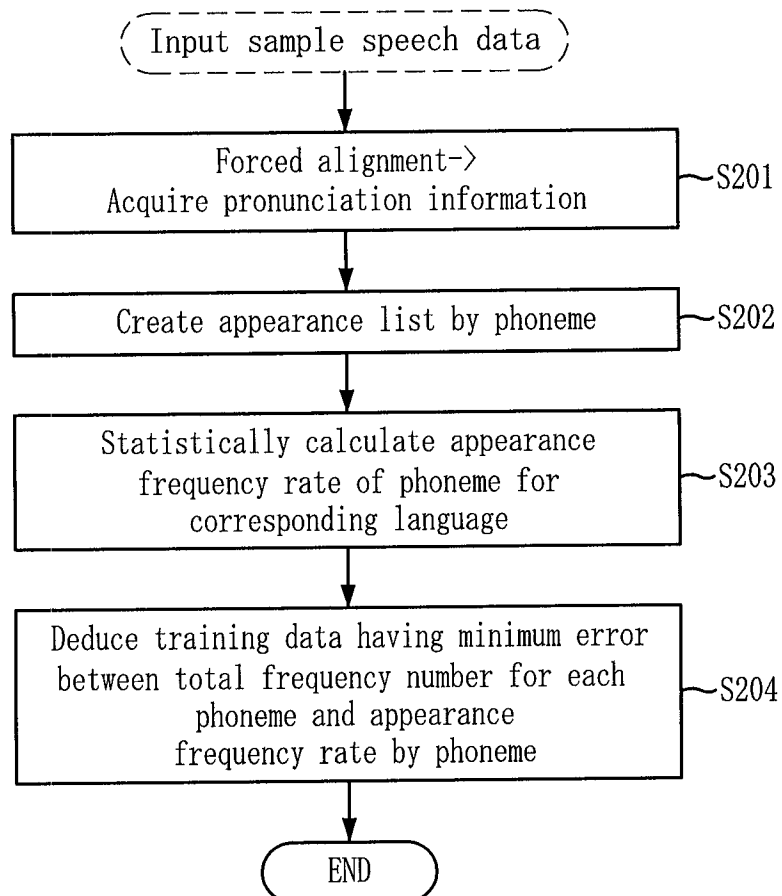
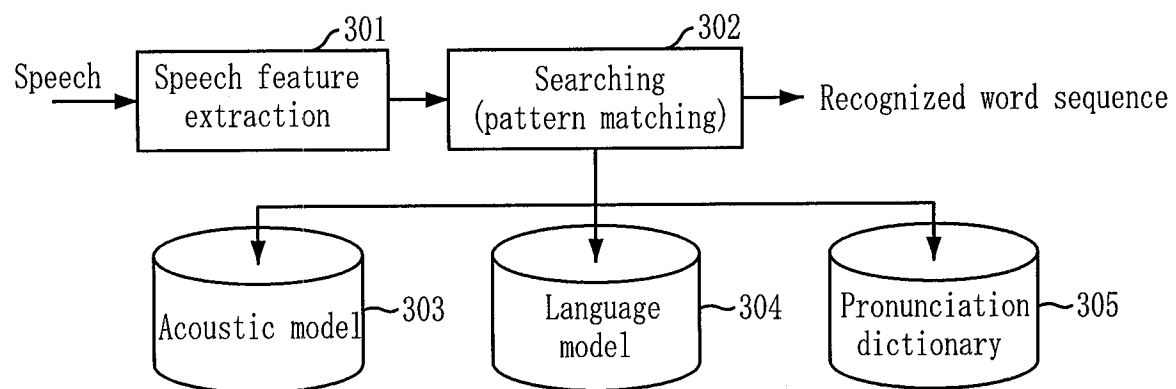


FIG. 2



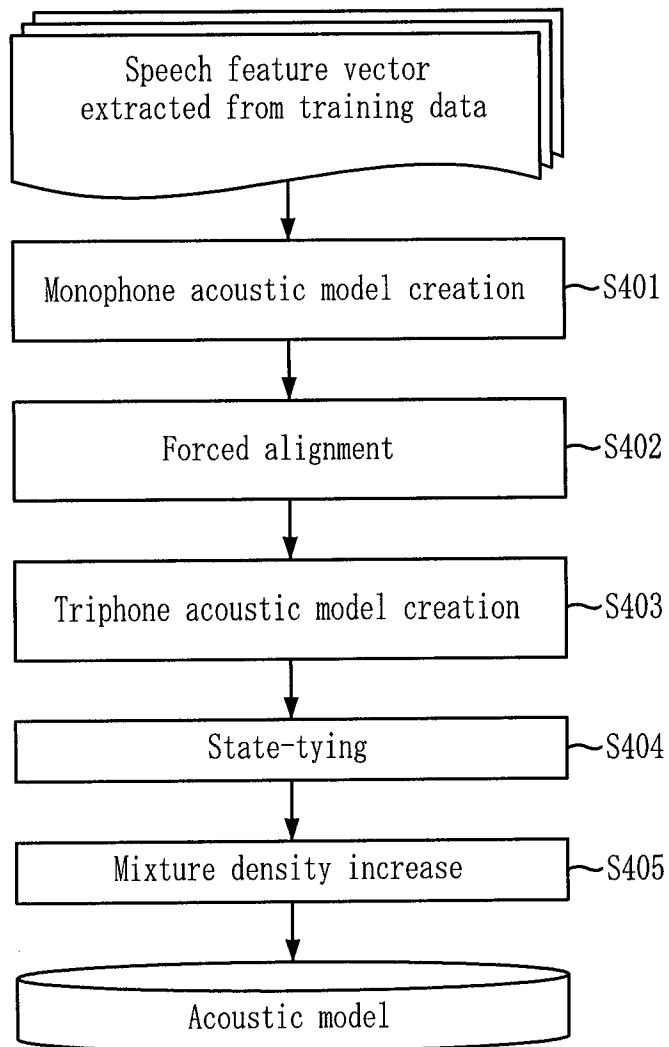
2/4

FIG. 3



3/4

FIG. 4



4/4

FIG. 5

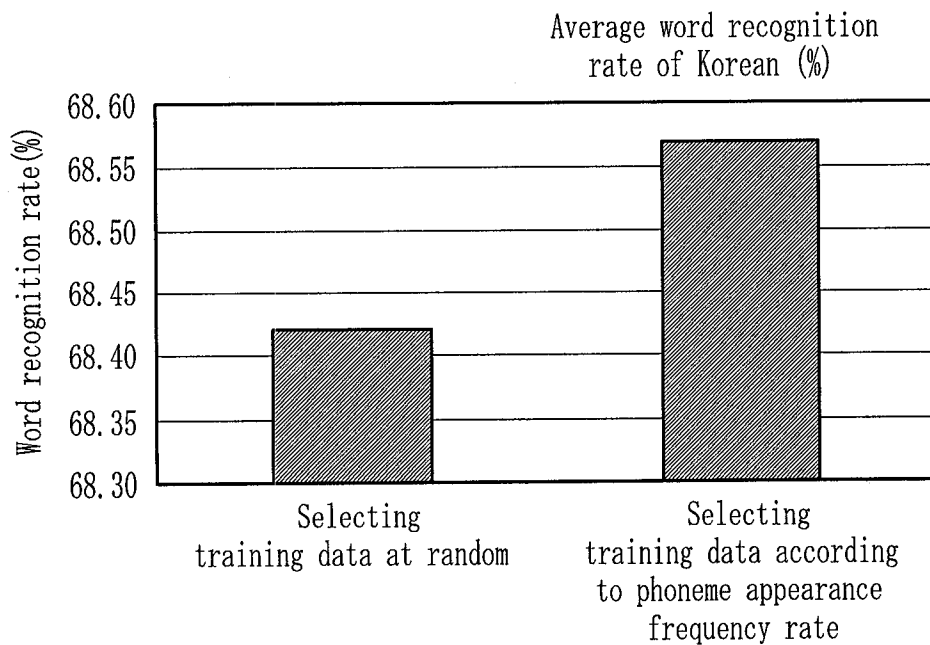
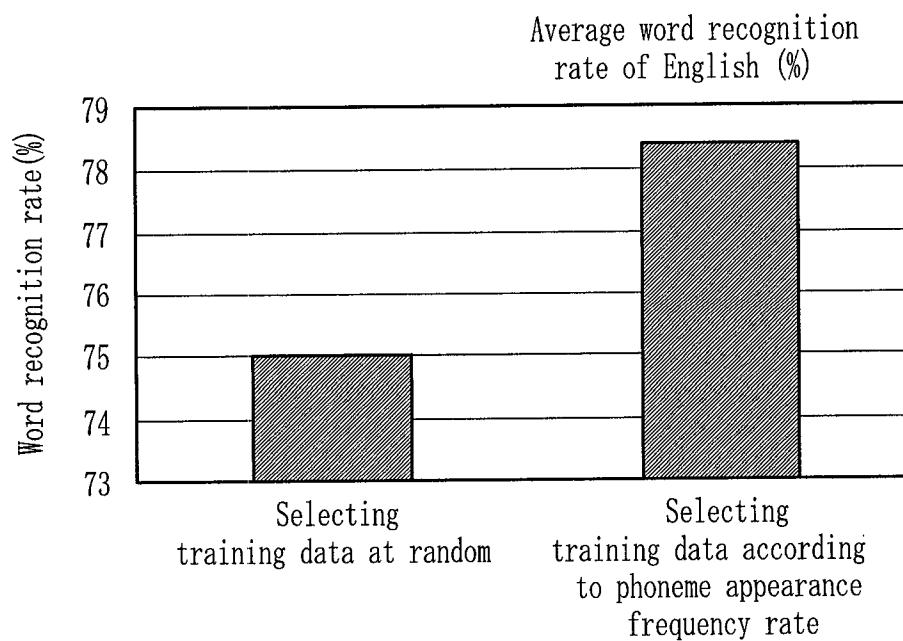


FIG. 6



INTERNATIONAL SEARCH REPORT

International application No.
PCT/KR2006/005830**A. CLASSIFICATION OF SUBJECT MATTER***G10L 15/06(2006.01)i, G10L 15/10(2006.01)i*

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)

IPC 8 : G10L 15/00 ~ G10L 15/28, G10L 11/00

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

Utility models and applications for Utility Models since 1975

Japanese Utility Models and application for Utility Models since 1975

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

eKIPASS(KIPO internal) "train*, phone*, prob*, list, forced, align*, voice, speech, data"

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	US 6182038 B1 (Sreeram B. and Stephen A.) 30 JANUARY 2001 See the abstract, figures 4-6,	1-10
A	JP H08-278794 A (SONY Co. Ltd.,) 22 OCTOBER 1996 See the abstract, figures 1,2,6,8,10	1-10
A	WO 2005/034086 A1 (ASAHI Co. Ltd.,) 14 APRIL 2005 See the abstract, figures 1,10,11,21,27	1-10
A	KR 10-1999-011915 A (LG Electronics Co. Ltd.,) 18 FEBRUARY 1999 See the abstract, page 2, figures 1,2	1-10
A	KR 10-2001-0004468 A (KT Co. Ltd.,) 15 JANUARY 2001 See the abstract, page 2-3, figures 2,3	1-10



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 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

11 APRIL 2007 (11.04.2007)

Date of mailing of the international search report

11 APRIL 2007 (11.04.2007)

Name and mailing address of the ISA/KR

Korean Intellectual Property Office
920 Dunsan-dong, Seo-gu, Daejeon 302-701,
Republic of Korea

Facsimile No. 82-42-472-7140

Authorized officer

KYUNG, Youn Jeong

Telephone No. 82-42-481-8536



INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No.

PCT/KR2006/005830

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
US06182038	30.01.2001	AT237176E	15.04.2003
		AU199914650A1	16.06.1999
		DE69813180C0	15.05.2003
		DE69813180T2	23.10.2003
		EP00954856B1	09.04.2003
		EP954856A1	10.11.1999
		FR2773413A1	09.07.1999
		FR2773413B1	19.05.2000
		GB2331826A	02.06.1999
		GB2331826B2	19.12.2001
		TW462037B	01.11.2001
		US6182038B1	30.01.2001
		W09928899A1	10.06.1999
JP08278794	22.10.1996	CN1140870A	22.01.1997
		DE69625950T2	24.12.2003
		EP0736857A2	09.10.1996
		EP0736857B1	29.01.2003
		JP3741156B2	01.02.2006
		KR1019960038734	21.11.1996
		US5848389A	08.12.1998
W02005034086A1	14.04.2005	EP1669979A1	14.06.2006
		EP1669979A4	27.12.2006
		KR2006087585A	02.08.2006
		US2005075875A1	07.04.2005
KR 10-1999-011915 A	18.02.1999	None	
KR 10-2001-004468 A	15.01.2001	None	