US006182037B1

(12) **United States Patent**
Maes

(10) **Patent No.: US 6,182,037 B1**
(45) **Date of Patent: *Jan. 30, 2001**

(54) **SPEAKER RECOGNITION OVER LARGE POPULATION WITH FAST AND DETAILED MATCHES**

(75) Inventor: **Stephane Herman Maes**, Danbury, CT (US)

(73) Assignee: **International Business Machines Corporation**

(*) Notice: This patent issued on a continued prosecution application filed under 37 CFR 1.53(d), and is subject to the twenty year patent term provisions of 35 U.S.C. 154(a)(2).

Under 35 U.S.C. 154(b), the term of this patent shall be extended for 0 days.

(21) Appl. No.: **08/851,982**

(22) Filed: **May 6, 1997**

(51) **Int. Cl.**[7] **G10L 17/00**
(52) **U.S. Cl.** **704/247**; 704/245
(58) **Field of Search** 704/246, 247, 704/250, 249, 243, 245, 244

(56) **References Cited**

U.S. PATENT DOCUMENTS

3,673,331	*	6/1972	Hair et al.	704/246
4,363,102		12/1982	Holmgren et al. .	
4,716,593	*	12/1987	Hirai et al.	704/247
4,720,863		1/1988	Li et al. .	
4,827,518		5/1989	Feustel et al. .	
4,947,436	*	8/1990	Greaves et al.	704/206
5,073,939		12/1991	Vensko et al. .	
5,121,428		6/1992	Uchiyama et al. .	
5,167,004		11/1992	Netsch et al. .	
5,189,727		2/1993	Guerreri .	
5,216,720		6/1993	Naik et al. .	
5,241,649		8/1993	Niyada .	
5,271,088		12/1993	Bahler .	
5,274,695		12/1993	Green .	
5,339,385		8/1994	Higgins .	
5,347,595	*	9/1994	Bokser	382/225
5,384,833		1/1995	Cameron .	
5,412,738		5/1995	Brunelli et al. .	
5,414,755		5/1995	Bahler et al. .	
5,522,012	*	5/1996	Mammone et al.	704/260
5,537,488	*	7/1996	Menon et al.	383/225

(List continued on next page.)

FOREIGN PATENT DOCUMENTS

59-111699	6/1984	(JP) .
61-2599	1/1986	(JP) .
4-15700	1/1992	(JP) .

OTHER PUBLICATIONS

T. Matsui et al.; "A Study of Model and a Priori Threshold Updating in Speaker Verification"; Technical Report of the Institute of Electronics, Information & Communications Engineers; SP95–120(1996–01); pp 21–26.

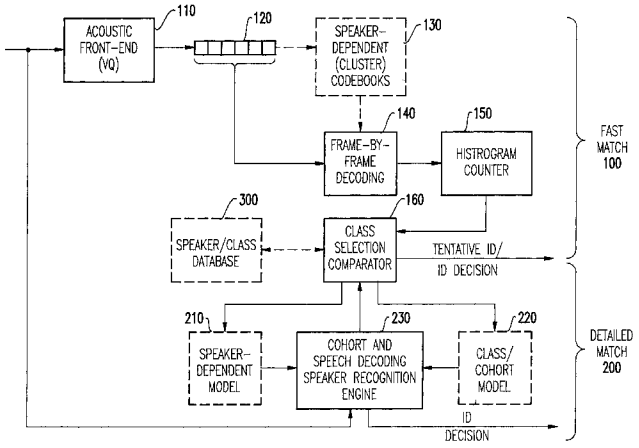
(List continued on next page.)

Primary Examiner—David R. Hudspeth
Assistant Examiner—Harold Zintel
(74) *Attorney, Agent, or Firm*—McGuireWoods, LLP; Paul J. Otterstedt

(57) **ABSTRACT**

Fast and detailed match techniques for speaker recognition are combined into a hybrid system in which speakers are associated in groups when potential confusion is detected between a speaker being enrolled and a previously enrolled speaker. Thus the detailed match techniques are invoked only at the potential onset of saturation of the fast match technique while the detailed match is facilitated by limitation of comparisons to the group and the development of speaker-dependent models which principally function to distinguish between members of a group rather than to more fully characterize each speaker. Thus storage and computational requirements are limited and fast and accurate speaker recognition can be extended over populations of speakers which would degrade or saturate fast match systems and degrade performance of detailed match systems.

23 Claims, 2 Drawing Sheets



U.S. PATENT DOCUMENTS

5,608,840	*	3/1997	Tsuboka	704/236
5,666,466	*	9/1997	Lin et al.	704/246
5,675,704	*	10/1997	Juang et al.	704/246
5,682,464	*	10/1997	Sejnoha	704/238
5,689,616	*	11/1997	Li	704/232
5,895,447	*	4/1999	Ittycheriah et al.	704/231

OTHER PUBLICATIONS

Parsons "Voice and Speech Processing" 1987, McGraw–Hill, pp. 332–336.*

Bahl et al "A fast approximate acoustic match for large vocabulary speech recognition" IEEE Transactions, Jan. 1993, pp. 59–67.*

Rudasi, Text–independent talker identification using recurrent neural networks: J Acoust Soc Am Supp 1 v 87, pg s104, 1990.*

"Merriam–Webster collegiate dictionary" pp. 211 and 550, 1993.*

Rabiner "Digital processing of speech signals" p. 478, 1978.*

Parsons "Voice and Speech Processing" 1987, McGraw–Hill, p. 175.*

Yu et al "Speakerrecognition using hidden Markov models, dynamic time warping and vector quantisation", Oct. 1995, IEEE, 313–318.*

Rosenberg, Lee, and Soone, Sub–Word Unit Talker Verification Using Hidden Markov Models, 1990, AT&T Bell Laboratories, pp 269–272.

Herbert Gish, Robust Discrimination in Automatic Speaker Identification, BBN Systems and Technologies Corporation, pp 289–292.

Naik, Netsch and Doddington, Speaker Verification Over Long Distance Telephone Lines, Texas Instruments Inc., pp 524–527.

* cited by examiner

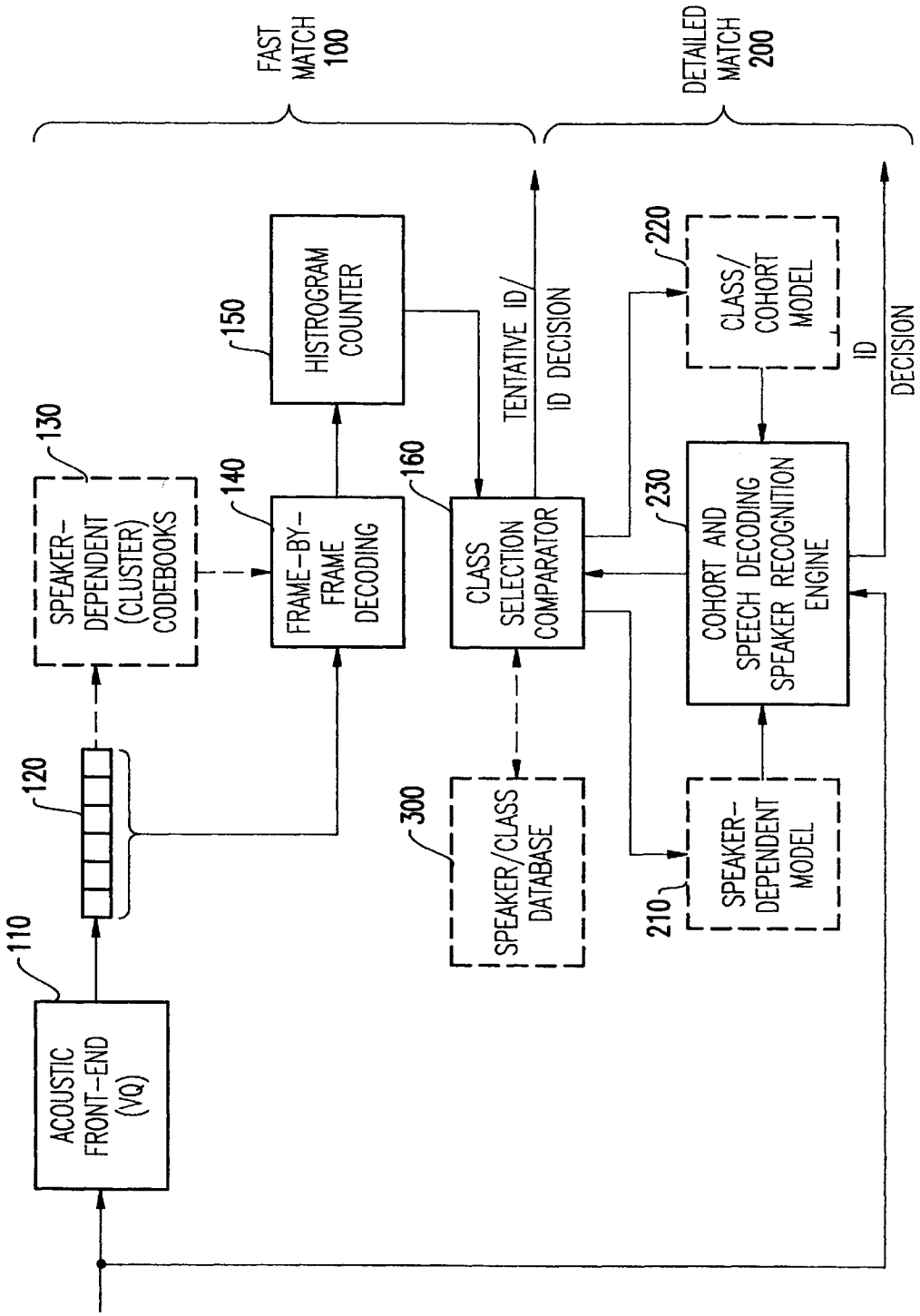


FIG. 1

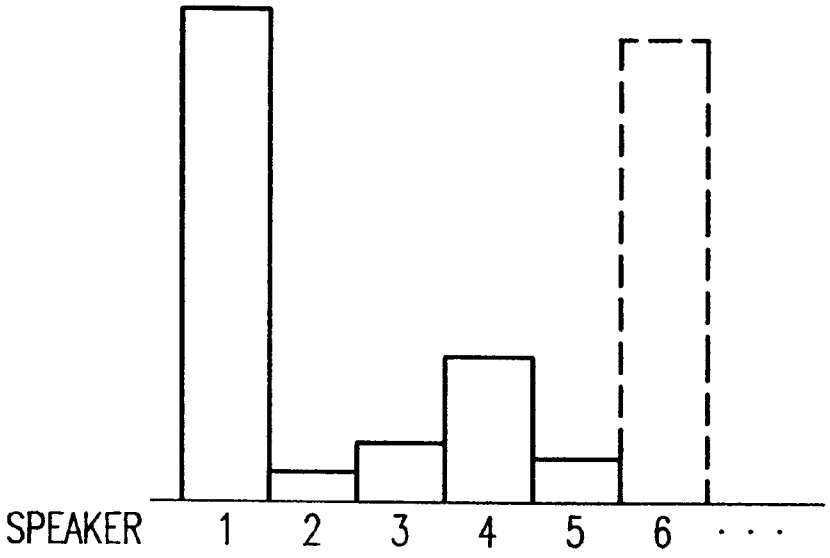


FIG.2A

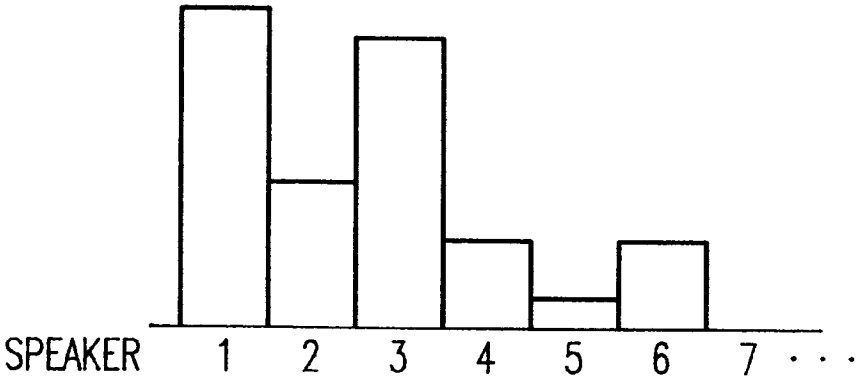


FIG.2B

SPEAKER RECOGNITION OVER LARGE POPULATION WITH FAST AND DETAILED MATCHES

BACKGROUND OF THE INVENTION

1. Field of the Invention

The present invention generally relates to speaker identification and verification in speech recognition systems and, more particularly, to rapid and text-independent speaker identification and verification over a large population of enrolled speakers.

2. Description of the Prior Art

Many electronic devices require input from a user in order to convey to the device particular information required to determine or perform a desired function or, in a trivially simple case, when a desired function is to be performed as would be indicated by, for example, activation of an on/off switch. When multiple different inputs are possible, a keyboard comprising an array of two or more switches has been the input device of choice in recent years.

However, keyboards of any type have inherent disadvantages. Most evidently, keyboards include a plurality of distributed actuable areas, each generally including moving parts subject to wear and damage and which must be sized to be actuated by a portion of the body unless a stylus or other separate mechanical expedient is employed. Accordingly, in many types of devices, such as input panels for security systems and electronic calculators, the size of the device is often determined by the dimensions of the keypad rather than the electronic contents of the housing. Additionally, numerous keystrokes may be required (e.g. to specify an operation, enter a security code, personal identification number (PIN), etc.) which slows operation and increases the possibility that erroneous actuation may occur. Therefore, use of a keyboard or other manually manipulated input structure requires action which is not optimally natural or expeditious for the user.

In an effort to provide a more naturally usable, convenient and rapid interface and to increase the capabilities thereof, numerous approaches to voice or sound detection and recognition systems have been proposed and implemented with some degree of success. Additionally, such systems could theoretically have the capability of matching utterances of a user against utterances of enrolled speakers for granting or denying access to resources of the device or system, identifying enrolled speakers or calling customized command libraries in accordance with speaker identity in a manner which may be relatively transparent and convenient to the user.

However, large systems including large resources are likely to have a large number of potential users and thus require massive amounts of storage and processing overhead to recognize speakers when the population of enrolled speakers becomes large. Saturation of the performance of speaker recognition systems will occur for simple and fast systems designed to quickly discriminate among different speakers when the size of the speaker population increases. Performance of most speaker-dependent (e.g. performing decoding of the utterance and aligning on the decoded script models such as hidden Markov models (HMM) adapted to the different speakers, the models presenting the highest likelihood of correct decoding identifying the speaker, and which may be text-dependent or text-independent) systems also degrades over large speaker populations but the tendency toward saturation and performance degradation is encountered over smaller populations with fast, simple

systems which discriminate between speakers based on smaller amounts of information and thus tend to return ambiguous results when data for larger populations results in smaller differences between instances of data.

As an illustration, text-independent systems such as frame-by-frame feature clustering and classification may be considered as a fast match technique for speaker or speaker class identification. However, the numbers of speaker classes and the number of speakers in each class that can be handled with practical amounts of processing overhead in acceptable response times is limited. (In other words, while frame-by-frame classifiers require relatively small amounts of data for each enrolled speaker and less processing time for limited numbers of speakers, their discrimination power is correspondingly limited and becomes severely compromised as the distinctiveness of the speaker models (each containing relatively less information than in speaker-dependent systems) is reduced by increasing numbers of models. It can be readily understood that any approach which seeks to reduce information (stored and/or processed) concerning speaker utterances may compromise the ability of the system to discriminate individual enrolled users as the population of users becomes large. At some size of the speaker population, the speaker recognition system or engine is no longer able to discriminate between some speakers. This condition is known as saturation.

On the other hand, more complex systems which use speaker dependent model-based decoders which are adapted to individual speakers to provide speaker recognition must run the models in parallel or sequentially to accomplish speaker recognition and therefore are extremely slow and require large amounts of memory and processor time. Additionally, such models are difficult to train and adapt since they typically require a large amount of data to form the model.

Some reduction in storage requirements has been achieved in template matching systems which are also text-dependent as well as speaker-dependent by reliance on particular utterances of each enrolled speaker which are specific to the speaker identification and/or verification function. However, such arrangements, by their nature, cannot be made transparent to the user; requiring a relatively lengthy enrollment and initial recognition (e.g. logon) procedure and more or less periodic interruption of use of the system for verification. Further and, perhaps, more importantly, such systems are more sensitive to variations of the utterances of each speaker ("intra-speaker" variations) such as may occur through aging, fatigue, illness, stress, prosody, psychological state and other conditions of each speaker.

More specifically, speaker-dependent speech recognizers build a model for each speaker during an enrollment phase of operation. Thereafter, a speaker and the utterance is recognized by the model which produces the largest likelihood or lowest error rate. Enough data is required to adapt each model to a unique speaker for all utterances to be recognized. For this reason, most speaker-dependent systems are also text-dependent and template matching is often used to reduce the amount of data to be stored in each model. Alternatively, systems using, for example, hidden Markov models (HMM) or similar statistical models usually involve the introduction of cohort models based on a group of speakers to be able to reject speakers which are too improbable.

Cohort models allow the introduction of confidence measures based on competing likelihoods of speaker identity and

DOCKET ALARM

Explore Litigation Insights

Docket Alarm provides insights to develop a more informed litigation strategy and the peace of mind of knowing you're on top of things.

Real-Time Litigation Alerts

Keep your litigation team up-to-date with **real-time alerts** and advanced team management tools built for the enterprise, all while greatly reducing PACER spend.

Our comprehensive service means we can handle Federal, State, and Administrative courts across the country.

Advanced Docket Research

With over 230 million records, Docket Alarm's cloud-native docket research platform finds what other services can't. Coverage includes Federal, State, plus PTAB, TTAB, ITC and NLRB decisions, all in one place.

Identify arguments that have been successful in the past with full text, pinpoint searching. Link to case law cited within any court document via Fastcase.

Analytics At Your Fingertips

Learn what happened the last time a particular judge, opposing counsel or company faced cases similar to yours.

Advanced out-of-the-box PTAB and TTAB analytics are always at your fingertips.

API

Docket Alarm offers a powerful API (application programming interface) to developers that want to integrate case filings into their apps.

LAW FIRMS
Build custom dashboards for your attorneys and clients with live data direct from the court.

Automate many repetitive legal tasks like conflict checks, document management, and marketing.

FINANCIAL INSTITUTIONS
Litigation and bankruptcy checks for companies and debtors.

E-DISCOVERY AND LEGAL VENDORS
Sync your system to PACER to automate legal marketing.