



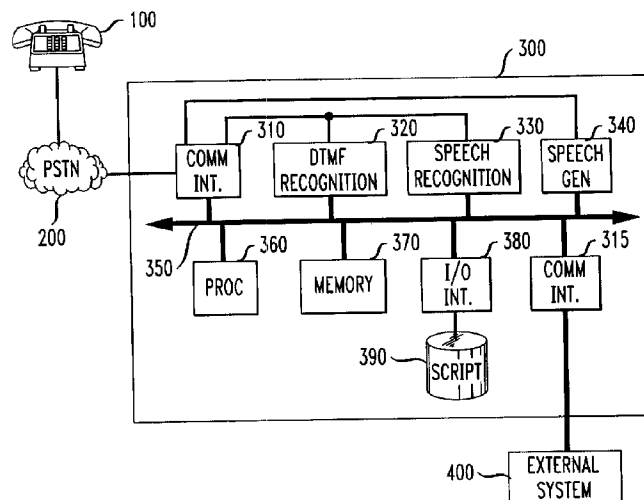
(72) BOYCE, Susan J., US  
(72) BROTMAN, Lynne S., US  
(72) BROWN, Deborah W., US  
(72) GOLDBERG, Randy G., US  
(72) HASZTO, Edward D., US  
(72) MARCUS, Stephen M., US  
(72) ROSINSKI, Richard R., US  
(72) WETZEL, William R., US  
(71) AT&T CORP., US

(51) Int.Cl.<sup>6</sup> G10L 9/18, G06F 17/20, G10L 9/00

(30) 1997/07/16 (08/895,298) US

(54) **RECONNAISSANCE DE LA PAROLE APPLIQUEE AUX  
COMMUNICATIONS TELEPHONQUES POUR FINS DE  
COLLECTE DE DONNEES**

(54) **TELEPHONE-BASED SPEECH RECOGNITION FOR DATA  
COLLECTION**



(57) Script opérationnel pour interactions entre des participants et un système de collecte automatique de données, créé par réglage d'un script expérimental au moyen d'essais successifs jusqu'à ce que le système atteigne un taux de reconnaissance au moins acceptable. Le système de collecte de données utilise une grammaire semi-contrainte, qui constitue une prise en charge pratique d'une gamme plus vaste d'entrées possibles qu'un menu. Ce système recueille des données par reconnaissance des énoncés formulés par les participants conformément au script opérationnel.

(57) A production script for interactions between participants and an automated data collection system is created by selectively tuning an experimental script through successive trials until a recognition rate of the system is at least an acceptability threshold. The data collection system uses a semi-constrained grammar, which is a practical accommodation of a larger range of possible inputs than a menu. The data collection system collects data by recognizing utterances from participants in accordance with the production script.



## **TELEPHONE-BASED SPEECH RECOGNITION FOR DATA COLLECTION**

### **ABSTRACT OF THE INVENTION**

A production script for interactions between participants and an automated data collection system is created by selectively tuning an experimental script through successive trials until a recognition rate of the system is at least an acceptability threshold. The data collection system uses a semi-constrained grammar, which is a practical accommodation of a larger range of possible inputs than a menu. The data collection system collects data by recognizing utterances from participants in accordance with the production script.

**TELEPHONE-BASED SPEECH RECOGNITION FOR DATA COLLECTION****BACKGROUND OF THE INVENTION**

The present invention relates to telephone-based data collection systems, and, more particularly, is directed to a voice recognition system using a script including prompts and corresponding recognition grammars which are empirically refined.

Many industries collect data from respondents by telephone. For qualitative information, a human agent is conventionally preferred. However, for certain applications, such as medical patient monitoring, dependence on human interviewers prohibitively increases the cost of gathering data as often as desired. For quantitative information, an automated data collection system is generally adequate, and in some cases may be preferred, e.g., people are more willing to give negative responses to an automated system than to a human agent. An automated system has the additional advantage relative to a human agent of eliminating a subsequent data entry step.

In one known service offering, customers call a special phone number and respond to pre-recorded questions using a telephone keypad. Limitations of this service offering include lack of universal availability of dual tone multi-frequency telephones (rotary telephones provide out-of-voiceband signals, and so are unsuitable for data entry) and an awkward interface due to availability of only twelve keys. Also, some users are uncomfortable with providing information by "typing".

## SUMMARY OF THE INVENTION

In accordance with an aspect of this invention, a system for collecting data is provided, comprising a memory for storing a production script including prompts and corresponding recognition grammars. The production script is created by selectively tuning an experimental script through successive trials until a recognition rate of the system is at least an acceptability threshold.

A speech generation circuit provides speech signals in accordance with the prompts of the production script. The speech signals are transmitted to a participant and utterances are received from the participant using a communication network interface. A speech recognition circuit recognizes the utterances from the participant in accordance with the recognition grammars.

It is not intended that the invention be summarized here in its entirety. Rather, further features, aspects and advantages of the invention are set forth in or are apparent from the following description and drawings.

## BRIEF DESCRIPTION OF THE DRAWINGS

Fig. 1 is a block diagram of a telephone-based data collection system;

Fig. 2 is a flowchart referred to in describing how the software for the data collection system of Fig. 1 is produced; and

Fig. 3 is a flowchart of the tuning process referenced in Fig. 2.

## 1 DETAILED DESCRIPTION OF THE PREFERRED EMBODIMENTS

2 Fig. 1 shows a typical environment for a telephone-based data collection  
3 system. A telephone set 100 having a standard key set (0, 1, ..., 9, \*, #) is used by a  
4 participant, such as from the participant's home or office. As used herein, a  
5 participant means a person providing information for collection to an automated data  
6 collection system.

7 Telecommunication network 200 comprises dial telephone lines and well-  
8 known switching equipment such as stored program control switches.

9 Data collection system 300 comprises communication interface 310 to network  
10 200, communication interface 315 to external system 400, dual tone multi-frequency  
11 (DTMF) signal recognition circuit 320, speech recognition circuit 330, speech  
12 generation/replay circuit 340, data bus 350, processor 360, memory 370, disk  
13 interface 380 and disk storage 390. Data collection system 300 is located within a  
14 network operated by a communication network operator. In other embodiments, data  
15 collection system 300 may be external to the network operated by the communication  
16 network operator.

17 Communication interface 310 functions to receive and place calls through  
18 network 200 under control of processor 360. Communication interface 310 supplies  
19 inbound call signals from network 200 to recognition circuits 320, 330, and receives  
20 outbound call signals from generation/replay circuit 340 for delivery to network 200.

21 Communication interface 315 functions to transmit data to and receive data  
22 from external system 400 under control of processor 360. In some embodiments,

1 communication interface 315 interfaces to PSTN 200, rather than a dedicated  
2 communication line (as shown).

3 DTMF signal recognition circuit 320 functions to recognize a DTMF signal  
4 produced by depression of a key on telephone set 100, to convert the DTMF signal to  
5 digital data and to supply the digital data to processor 360 via data bus 350.

6 Speech recognition circuit 330 is adapted to receive a speech signal spoken by  
7 a participant to telephone set 100, to process the speech signal to obtain a word string  
8 or determination that the speech signal is unrecognizable, and to deliver its processing  
9 results to processor 360 via data bus 350.

10 Speech generation/replay circuit 340 is adapted to receive a control signal  
11 from processor 360 via data bus 350, and in response thereto, to generate a speech  
12 signal or to replay a pre-recorded speech signal, and to provide the generated or  
13 replayed speech signal to communications interface 310.

14 Processor 360 serves to control circuits 310, 320, 330 and 340 to collect  
15 information from participants in accordance with a predetermined script including  
16 prompts and corresponding recognition grammars, to store the collected information  
17 and to make the stored information available to a recipient of the information.

18 Processor 360 is operative in conjunction with memory 370, disk interface 380 and  
19 disk storage 390 in a conventional manner which is well-known to one of ordinary  
20 skill in the art. It will be appreciated that the predetermined script is stored on disk  
21 storage 390. At initialization of data collection system 300, typically when power is  
22 supplied thereto, appropriate portions of the predetermined script are loaded into

1 memory 370.

2 An important feature of the present invention is the ability to provide a semi-  
3 constrained vocabulary recognition response grammar. That is, rather than requiring  
4 a caller to enter data via manual entry on a telephone keypad, a data collection system  
5 according to the present invention supports very flexible speech recognition. The  
6 data collection system does not require "training" for each caller, but rather can  
7 accommodate new callers. The data collection system is not limited to a small  
8 recognition vocabulary, but rather accommodates a conversational style dialog with a  
9 caller. However, the recognition grammar is semi-constrained in accordance with an  
10 empirical grammar construction process, described in detail below.

11 To support a larger participant community, various elements of the system  
12 shown in Fig. 1 may be replicated. For example, to support a plurality of  
13 simultaneous calls, multiple occurrences of communication interface 310, DTMF  
14 signal recognition circuit 320, speech recognition circuit 330, and speech  
15 generation/replay circuit 340 may be provided.

16 Typical applications for the data collection system 300 include collection of  
17 medical data from patients; collection of purchase information from customers, e.g.,  
18 to make a rental car reservation or to order merchandise or services; and collection of  
19 survey data from subjects.

20 Initially, the recipient of information (such as a hospital or company testing a  
21 new drug) sets up a script for the service dialog, and registers participants (such as  
22 doctors and patients) possibly including a voice print for verification. For

1 applications involving confirmation of the caller's identity, multiple levels of  
2 verification may be elected by the recipient of information, such as tiers of  
3 verification.

4 On a regular or as needed basis, patients call in to provide information, such  
5 as:

6  
7 System How are you feeling on a scale of one to five, with one being  
8 very good and five being very poor?

9  
10 Patient Ummm ... three.

11  
12 System You're feeling fair. Is that correct?

13  
14 Patient Yes.

15  
16 Generally, a participant employs telephone set 100 to dial a call through  
17 network 200 to data collection system 300 which is configured to receive the  
18 participant's call and engage the participant in a speech based dialog to collect mostly  
19 spoken information. Using automated speech recognition, data collection system 300  
20 converts the spoken information into response data, which is verified with the speaker  
21 via voice synthesis or pre-recorded segments. The speech recognition supports "word  
22 spotting" (the desired response is recognized from a longer utterance) and "barge in"



(the speaker utters something before the system has completed its spoken question or instruction). An example of word spotting is processing "Ummm ... three" to extract the response "three".

When data collection system 300 is used to interview an out-patient at a mental hospital, a prompt may be:

System        Please state all of the drugs that you are presently taking.

The semi-constrained vocabulary recognition response grammar is all of the drugs prescribed by the mental hospital. An example grammar is:

ativan	lithium	stelazine
buspar	luvox	symmetryl
clozaril	paxil	thorazine
cogentin	prolixin	tofranil
depakote	prozac	xanax
haldol	risperidol	zoloft
klonopin	ritalin	zyprexa

It will be appreciated that, as the list of allowable responses grows, is it much more convenient for the participant to respond to an open-ended question which has a semi-constrained response vocabulary than for the participant to listen to a menu and enter

1 a selection, possibly with a delimiter. An open-ended question is particularly  
2 convenient when the participant is being asked about information which they know,  
3 e.g., personal statistics or desired selection, and thus for which they do not require a  
4 menu.

5 In some embodiments, data collection system 300 also includes speaker  
6 verification based on voice prints, which entails registration of participants including  
7 provision of a baseline voice print; confirmation of caller identity, such as by  
8 provision of prerecorded distinguishing information (e.g., mother's maiden name);  
9 call transfer, e.g., to voice mail services or a person; and outbound calling to  
10 participants to collect information, or as a reminder to call the automated system and  
11 provide information.

12 The collected response data are disseminated as printed reports, possibly  
13 transmitted via facsimile or as digital data. In some embodiments, the collected  
14 response data are made available through a voice based interface tailored for the  
15 recipient of information, e.g., the recipient calls data collection system 300 and  
16 queries the status of selected participants and/or newly collected information. In  
17 some embodiments, data collection system 300 processes the collected data, e.g.,  
18 detects predetermined events, or makes the collected data available to a separate data  
19 processing system (not shown).

20 External system 400 may be used to supply data to data collection system 300,  
21 e.g., a daily update of scheduling information or real-time reservation availability.  
22 External system 400 may additionally or alternatively be a destination for information

1 collected by data collection system 300.

2 Fig. 2 is a flowchart showing how the predetermined script including prompts  
3 and corresponding recognition grammars used by data collection system 300 is  
4 produced.

5 At step 500, an experimental script is prepared. Typically, producers of the  
6 script provide their best guesses as to (i) the sequence and phrasing of questions most  
7 likely to elicit desired information from participants, that is, the prompts, and (ii) the  
8 types of responses which participants are most likely to give. Based on the most  
9 likely responses, a recognition grammar is created. If a response is not understood,  
10 provision of a question to elicit understandable information, referred to as re-  
11 prompting, may be included as part of the script. How extensively re-prompting is  
12 employed depends on the criticality of the information and the motivation of the  
13 participant, e.g., a patient versus an anonymous caller.

14 At step 510, a trial of the experimental script is conducted with a test group of  
15 participants. The participants' responses are recorded.

16 At step 520, the results of the trial are analyzed, and an overall recognition  
17 rate is obtained. In most cases, selected questions exhibit a low recognition rate,  
18 while other questions have an acceptable recognition rate.

19 The recognition rate of the data collection system 300 is defined as follows:

$$\text{Recognition Rate} = \frac{\text{Correct Recognitions}}{\text{Total Attempted Recognitions}}$$

where

Correct Recognitions = IN\_ACCEPT + OUT\_REJECT

Total Attempted Recognitions = Correct Recognitions + IN\_REJECT +  
OUT\_ACCEPT

IN\_ACCEPT = responses in grammar that were accepted

IN\_REJECT = responses in grammar that were rejected

OUT\_ACCEPT = responses out of grammar that were accepted

OUT\_REJECT = responses out of grammar that were rejected

At step 530, tuning is performed to improve the recognition rate. Tuning is discussed in detail below. Most of the tuning is performed manually. The result of step 530 is a revised experimental script. Tuning is an important aspect of creation of a semi-constrained grammar.

At step 540, it is determined whether the revised experimental script is appropriate for batch testing. Generally, if the changes made during tuning do not affect the prompts or prompt sequence, then batch testing is appropriate.

If batch testing is appropriate, at step 550, a batch test is conducted. Batch testing is rerunning collected data with a revised speech recognition system, and is

1 generally faster and more economical than another user trial.

2 If, at step 540, it is determined that batch testing is inappropriate, step 510 is  
3 repeated using the revised experimental script. In most cases, user trials are  
4 conducted with a test group having participants other than those who participated in a  
5 previous user trial.

6 Step 520 is repeated on the results of the most recent user testing or batch  
7 trial. If the recognition rate is unacceptable, the tuning, testing, trial and acceptability  
8 steps 530, 540, 510 and 520, respectively, are repeated until the recognition rate  
9 attains an acceptable threshold.

10 When an acceptable recognition rate is attained, the experimental script used  
11 in the associated trial is defined to be the production script, and, at step 560, the  
12 process of producing the predefined (or production) script is complete.

13 Fig. 3 is a flowchart of the tuning process shown in step 530 of Fig. 2. Much  
14 of the tuning process is performed manually, and the order in which the various  
15 tuning techniques are applied varies from artisan to artisan. The tuning techniques  
16 typically used include tuning the recognition grammar, tuning prompts, modifying re-  
17 prompts, pruning the grammar, tuning recognition parameters, pre- and post-  
18 processor tuning, and tuning the sequence of prompts. Other tuning techniques may  
19 also be used.

20 *Tuning the recognition grammar* refers to expanding the recognition grammar  
21 to include what people say, for example, a larger vocabulary of substantive responses,  
22 surrounding prepositions, and surrounding phrases.

1       *Pruning the grammar* refers to excluding rarely used words from the  
2 recognition grammar. Interestingly, although removing rarely used words ensures a  
3 recognition error when a participant says the removed word, the performance  
4 improvement in misrecognitions, that is, the reduction in recognition errors for more  
5 frequently said words which are misrecognized as one of the rarely used words, can  
6 be larger than the performance degradation due to the ensured errors.

7       *Tuning recognition parameters* refers to adjusting the parameters of the  
8 software used in speech recognition circuit 330. Recognition software is available  
9 from various vendors such as the HARK™ Telephony Recognizer, Release 3.0, from  
10 Bolt, Beranek and Newman, Cambridge, Massachusetts 02138, described in  
11 Document No. 300-3.0 (July 1995), the disclosure of which is hereby incorporated by  
12 reference. An example of a recognition parameter is the "rejection parameter", which  
13 attempts to maximize the "proper rejections" curve and minimize the "false  
14 rejections" curve, which typically have different shapes. The rejection parameter is  
15 usually computed to maximize the following expression:

$$(proper\ rejections) - (false\ rejections)$$

$$= OUT\_REJECT - IN\_REJECT$$

18 Another example of a recognition parameter is speaker gender sensitivity.

19       *Tuning prompts* refers to changing a prompt. An example is provided below  
20 (changing PROMPT 2 of the experimental script discussed below).

21       *Modifying re-prompts* refers to adding re-prompts or altering re-prompts,  
22 which are used when the recognition system fails to recognize a response. An

1 example is provided below (adding PROMPT R1 of the experimental script discussed  
2 below).

3 *Tuning the sequence of prompts* refers to changing the order in which prompts  
4 are presented. The purpose is to provide a better context for the participant, thereby  
5 avoiding errors due to participant confusion.

6 *Pre- and post-processor tuning* refers to altering how external information is  
7 employed by speech recognition circuit 330 and processor 360. An example of  
8 external information is an account number or a confusion matrix. Processor 360 may  
9 use the output of speech recognition circuit 330 to enhance account number  
10 recognition.

11 An example of post-processor tuning is using a confusion matrix to improve  
12 the recognition accuracy. Use of confusion matrices with account numbers is  
13 described in application serial no. 08/763,382 (attorney docket no. BROWN 1-3-1-3),  
14 the disclosure of which is hereby incorporated by reference.

15 An example of pre-processor tuning is constraining the account numbers which  
16 may exist to improve recognition accuracy. Choosing alpha-numerics to improve  
17 recognition accuracy is described in application serial no. 08/771,356 (attorney docket  
18 no. BROWN 2-4-1-4-3), the disclosure of which is hereby incorporated by reference.

19 At step 600 of Fig. 3, the results of a user trial conducted at step 510 of Fig. 2  
20 are analyzed.

21 If, at step 610, it is determined that a sufficient percentage of responses were  
22 out of grammar, then at step 620, the out of grammar responses are analyzed to

1 determine whether they are clustered or grouped. A sufficient percentage may be, for  
2 example, at least 10% of the responses. If the responses are clustered, at step 622,  
3 the recognition grammar is augmented (expanded) to include what people actually  
4 said, and a batch test can be performed to evaluate the results of the grammar tuning.

5 Step 622 can be performed entirely by a computer programmed appropriately. If the  
6 responses are not clustered, then at step 625, the prompts are tuned, and at step 627  
7 the re-prompts are adjusted. After modifying the prompts or re-prompts, a batch test  
8 is inappropriate, so a user trial should be performed.

9 If, at step 610, it was determined that an out of grammar situation was not  
10 present, then at step 630, it is determined whether the in-grammar recognition  
11 accuracy is low, for example, less than 80% correct recognition of in grammar  
12 responses. If the in-grammar recognition accuracy is low, then at step 632, the  
13 grammar is pruned and at step 634, the recognition parameters are tuned. Step 632  
14 can be performed entirely by a computer programmed appropriately. A batch test  
15 may be performed to evaluate the results of the grammar and parameter tuning.

16 If, at step 630, it is determined that the in-grammar recognition accuracy is  
17 acceptable, then at step 640, it is determined whether the responses indicate that  
18 people are getting confused, for example, by giving wholly inappropriate responses to  
19 certain prompts. If people are getting confused, then, at step 642, the prompt  
20 sequence is modified, and the prompts and re-prompts may also be adjusted.

21 If, at step 640, it is determined that people are not confused, i.e., their out-of-  
22 grammar responses are not clustered, then at step 645, pre-processor tuning and post-



processor tuning are performed. A batch test may be performed to evaluate the results of the pre-processor and post-processor tuning.

An example of production of a script will now be discussed. The application is a system for collecting information for car rental reservations.

An experimental script included the following prompts and corresponding grammars, where triangular brackets ( < > ) indicate that a word in the indicated category is expected:

PROMPT 1: Where will you be renting?

The recognition grammar for PROMPT 1 was

< AIRPORT >

< CITY > < STATE >

< CITY > < STATE > < AIRPORT >

PROMPT 2: Picking up and returning back to < LOCATION > ?

The recognition grammar for PROMPT 2 was:

YES

NO

A first trial was conducted. The recognition rate was about 70%, that is, 30% of the participants said things that could not be processed by the telephone data collection system. This recognition rate was unacceptable. Tuning was performed, as described below.

The responses to PROMPT 1 included:

1 Chicago  
2 from O-Hare  
3 O-Hare  
4 Chicago O-Hare  
5 O-Hare Airport  
6 Chicago O-Hare Airport  
7  
8 New York  
9 JFK Airport  
10 at JFK  
11 from Newark Airport  
12 uh landing um at Newark  
13 Kennedy International  
14  
15 Los Angeles  
16 L-A-X Airport  
17 L-A-X  
18 LA International  
19

20 Participants used many variations on place names, occasionally added prepositions to  
21 their responses, and provided extraneous utterances such as "uh", "ah", "um", "er"  
22 and so on. The grammar was adjusted to account for this.

1 The responses to PROMPT 2 were quite varied, so the prompt was eliminated,  
2 and the following prompts used in place thereof:

3 PROMPT 2A: Are you returning to the same location?

4 PROMPT 2B: Where will you be returning the car?

5 Additionally, re-prompting capability was added, including:

6 PROMPT R1: I'm sorry, I didn't understand what you said.

7 If a complete sentence was provided as an answer, re-prompting was considered faster  
8 than trying to extract information from the sentence.

9 For ambiguous airports and cities, additional prompting was added:

10 PROMPT 1A: Is that Charleston, South Carolina or Charleston,  
11 West Virginia?

12 PROMPT 1B: Is that Boston, Massachusetts or Austin, Texas?

13 PROMPT 1C: Is that Columbus, Ohio, Columbus, Mississippi  
14 or Columbus, Georgia?

15 PROMPT 1D: Is that Portland, Oregon or Portland, Maine?

16 PROMPT 1E: Is that Washington, National, Dulles or BWI?

17 PROMPT 1F: Is that Dulles International or Dallas, Texas?

18 PROMPT 1G: Is that JFK, Lagueardia or Newark International?

19 Also, at any point, if the participant said "help", an appropriate response was  
20 provided.

21 A second trial was conducted, and the responses were recorded. The  
22 recognition rate was about 90%, that is, 10% of the participants said things that could

1 not be processed by the telephone data collection system. This recognition rate was  
2 unacceptable. Tuning was performed. The speech models used in the recognizer  
3 were changed to include large country-wide speech data, that is, more regional  
4 pronunciations were accommodated. Also, the speech models were changed to  
5 incorporate better phoneme recognition. Better rejection was added to the grammar.  
6 For example, if a response of "um" or "er" was provided, the system repeated the  
7 question.

8 A third trial was conducted, in which the pre-recorded responses of the second  
9 trial were played back; no new participant data was collected. Results were  
10 encouraging.

11 A fourth trial was conducted, with live participants. The recognition rate was  
12 about 98%. This was considered acceptable, that is, the acceptability threshold was  
13 no more than 98%. The script used in the fourth trial was defined to be the  
14 production script.

15 Although an illustrative embodiment of the present invention, and various  
16 modifications thereof, have been described in detail herein with reference to the  
17 accompanying drawings, it is to be understood that the invention is not limited to this  
18 precise embodiment and the described modifications, and that various changes and  
19 further modifications may be effected therein by one skilled in the art without  
20 departing from the scope or spirit of the invention as defined in the appended claims.

What is claimed is:

1. A system for collecting data, comprising:
  - a memory for storing a production script including prompts and corresponding recognition grammars;
  - a speech generation circuit for providing speech signals in accordance with the prompts of the production script;
  - a communication network interface for transmitting the speech signals to a participant and for receiving utterances from the participant; and
  - a speech recognition circuit for recognizing the utterances from the participant in accordance with the recognition grammars;wherein the production script is created by selectively tuning an experimental script through successive trials until a recognition rate of the system is at least an acceptability threshold.
2. The system of claim 1, further comprising a data storage device for storing the recognized utterances from the participant.
3. The system of claim 1, wherein the utterances of the participant are provided to the automated data collection system via telephone.
4. The system of claim 1, wherein the automated data collection system is located within a network operated by a communication network operator.
5. The system of claim 1, wherein the automated data collection system interacts with an external system to recognize utterances.
6. The system of claim 1, wherein at least one of the recognition grammars includes a semi-constrained vocabulary recognition response grammar.

FIG. 1

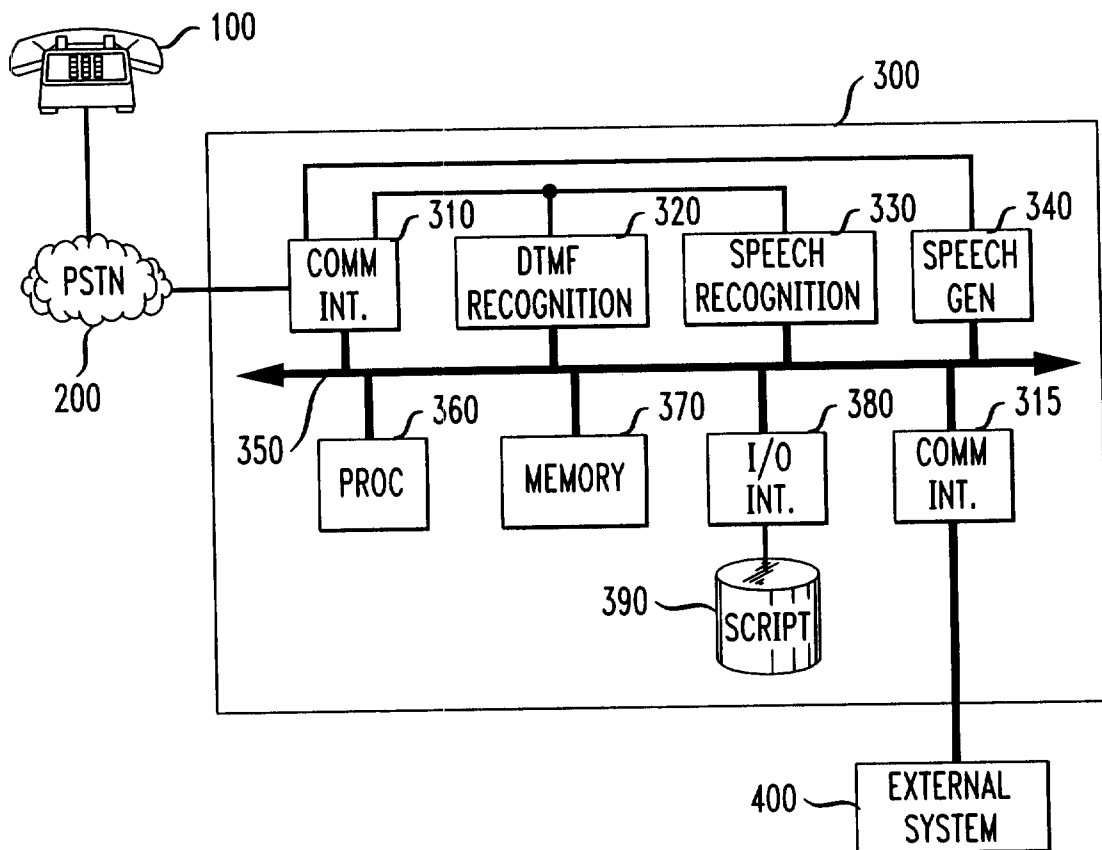


FIG. 2

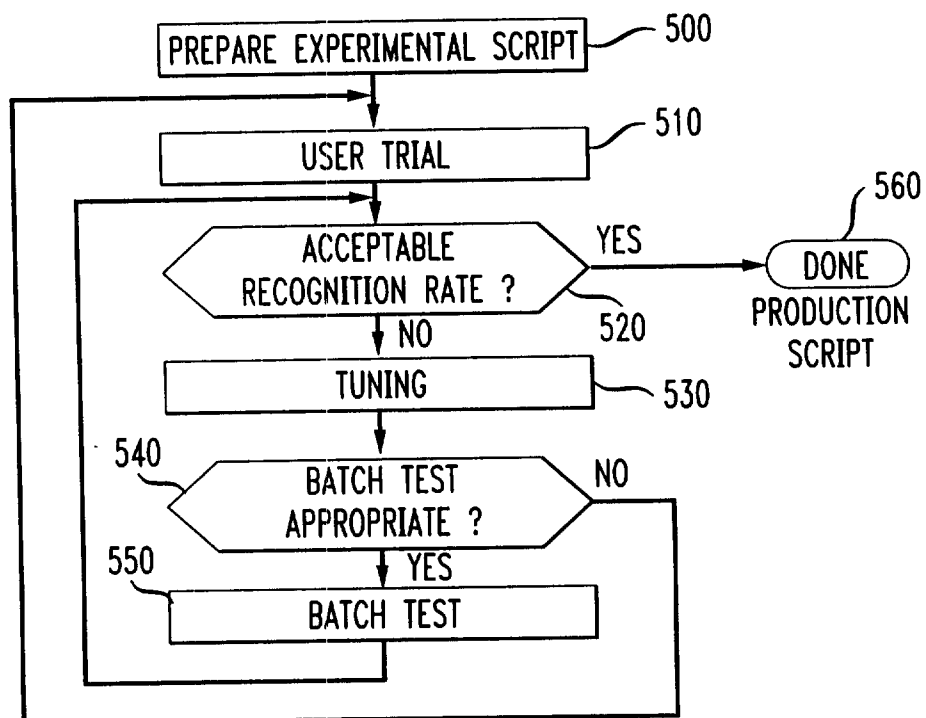


FIG. 3

