**Title:** EFFICIENTLY REDUCING TRANSCRIPTION ERROR USING HYBRID VOICE TRANSCRIPTION

**Abstract:** A system (10) and method (40) for efficiently reducing transcription error using hybrid voice transcription is provided. A voice stream (22) is parsed into utterances. An initial transcribed value and corresponding recognition score (113) are assigned to each utterance. A transcribed message (23) is generated and includes the initial transcribed values. A threshold is applied to the recognition scores (113) to identify those utterances with recognition scores (113) below the threshold as questionable utterances (111). At least one questionable utterance (111) is compared to other questionable utterances (111) from other calls and a group of similar questionable utterances (111) is formed. One or more of the similar questionable utterances (111) is selected from the group. A common manual transcription value (124) is received for the selected similar questionable utterances (111). The common manual transcription value (124) is assigned to the remaining similar questionable utterances (111) in the group.

---

**Fig. 9.**

![Image of a graph with transcribed values and thresholds]
before the expiration of the time limit for amending the claims and to be republished in the event of receipt of amendments (Rule 48.2(h))
EFFICIENTLY REDUCING TRANSCRIPTION ERROR USING HYBRID VOICE
TRANSCRIPTION

TECHNICAL FIELD

The present invention relates in general to speech recognition, in particular, to a system and method for efficiently reducing transcription error using hybrid voice transcription.

BACKGROUND ART

Automated speech recognition is commonly used in call centers to convert voice signals from callers into text. Generally, a voice or voice recording is received into the call center and speech is obtained. The speech is input into an automated speech recognition system, which parses the speech into short segments and assigns phonemes to the segments. The phonemes are analyzed and compared to a grammar of known words, phrases, and sentences to provide text values for the speech.

Once converted, the text can be used to store a record of a call, to identify characteristics of the call, or as a confirmation of the call. Speech recognition is also widely used in other fields, including the legal field for court reporting and dictation, and the medical field. The benefits of automated speech recognition include a reduction in the cost of employees required to manually transcribe voice messages and an increase in transcription speed. However, a lack of transcription accuracy is a barrier to widespread use of automated speech recognition.

A conventional approach using automated speech recognition and manual transcription has been implemented as an attempt to address and improve transcription accuracy. Generally, a voice message is first transcribed via automated speech recognition. Subsequently, an accuracy threshold is applied to the transcribed voice message. If the accuracy of the transcribed message is above the threshold, the transcribed voice message is provided to a user or stored. Whereas, if the accuracy of the transcribed message is below the threshold, the entire voice message is transmitted to a human transcriber for manual transcription. During manual transcription, each voice message transmitted to the human transcriber is separately processed, which can be expensive and time-consuming.

In large call centers, hundreds or thousands of calls can be received within a relatively short time period. During this time period, common utterances are received into the call center from different callers as voice. According to the conventional approach described above, if the transcription of the voice message fails to meet a threshold accuracy, the entire voice message is processed, which can be expensive and time-consuming.
then manually transcribed, which can be costly and time consuming. Thus, the conventional
approach fails to reduce error by identifying similar utterances during a specified time period,
manually transcribing at least one of the utterances, and then assigning the transcribed value to
the remaining similar utterances.

Therefore, there is a need for providing efficient and cost effective approaches for
reducing transcription error via a hybrid of automatic transcription and manual transcription.

A system and method for reducing transcription error is provided. A voice stream, such
as a live voice stream or a recording, is collected and parsed into speech utterances. The voice
stream is transcribed by assigning a transcription value and confidence score to each utterance.

A threshold is applied to the confidence scores and only those utterances with confidence scores
below the threshold are selected as questionable utterances. The questionable utterances have a
higher likelihood of being associated with an incorrect transcription value and should be further
analyzed. One of the questionable utterances is selected and a pool of similar questionable
utterances from other voice streams is generated. Subsequently, a sample is selected from the
pool, and each utterance in the sample is manually transcribed. If a common transcription value is
assigned to each utterance in the sample, the common transcription value is then assigned to the
remaining questionable utterances in the pool and incorporated into respective transcription
messages. Otherwise, the remaining questionable utterances are each manually transcribed.

One embodiment provides a system and method for efficiently reducing transcription
error using hybrid voice transcription is provided. A voice stream is parsed from a call into
utterances. A transcribed message is generated for the call and includes the initial transcribed
values. A threshold is applied to the recognition scores to identify those utterances with
values. A threshold is applied to the recognition scores to identify those utterances with
similar questionable utterances in the group. One or more of the similar questionable utterances
are selected from the group. A common manual transcription value is assigned to the selected similar
questionable utterances. The common manual transcription value is assigned to the remaining
similar questionable utterances in the group.
A further embodiment provides a system and method for hybrid voice transcription error reduction. A call is parsed into speech utterances, and a transcribed message is generated for the call by assigning an initial transcribed value and corresponding recognition score to each utterance. A confidence threshold is applied to the recognition scores. Those utterances with initial transcribed values that fall below the threshold are selected as questionable utterances. A sample is generated for at least one of the questionable utterances. Other questionable utterances are identified from other calls that are similar to the at least one questionable utterance. The at least one questionable utterance and similar questionable utterances are grouped as related utterances. One or more of the related utterances are selected as the sample. A common manual transcription value is received for each of the selected related utterances in the sample and is assigned to the remaining related utterances in the group, the remaining related utterances in the group.

Still other embodiments of the present invention will become readily apparent to those skilled in the art from the following detailed description, wherein is described embodiments of the invention by way of illustrating the best mode contemplated for carrying out the invention. As will be realized, the invention is capable of other and different embodiments and its several details are capable of modifications in various obvious respects, all without departing from the spirit and the scope of the present invention. The drawings and detailed description are to be regarded as illustrative in nature and not as restrictive.

**DESCRIPTION OF THE DRAWINGS**

FIGURE 1 is a functional block diagram showing a system for hybrid voice transcription, in accordance with one embodiment.

FIGURE 2 is a functional block diagram showing, by way of example, the hybrid voice transcription system of FIGURE 1 as incorporated into a call center.

FIGURE 3 is a process flow diagram showing a method for hybrid voice transcription, in accordance with one embodiment.

FIGURE 4 is a process flow diagram showing, by way of example, a method for building pools of similar utterances for use in the process of FIGURE 3.

FIGURE 5 is data flow diagram showing, by way of example, similarity metrics for use in the processes of FIGURE 4.

FIGURE 6 is a data flow diagram showing, by way of example, metrics for obtaining a sample from the pools of FIGURE 4.

FIGURE 7 is a block diagram showing, by way of example, an initial transcribed message.
FIGURE 8 is a block diagram showing, by way of example, the transcribed message of FIGURE 7 with identification values representing transcribed values for questionable utterances. FIGURE 9 is a block diagram showing, by way of example, a pool of similar questionable utterances. FIGURE 10 is a block diagram showing, by way of example, a transcribed utterance value chart.

**BEST MODE FOR CARRYING OUT THE INVENTION**

Automated speech recognition provides fairly quick and convenient transcription of voice to text that is frequently used in call centers, as well as the legal and medical industries to reduce costs and increase efficiency. However, conventional systems for automated speech recognition often times achieve the ability to reduce costs and increase efficiency at the expense of transcription accuracy. Quick, easy, and accurate reduction of any transcription errors is desirable during transcription. Reducing transcription errors using hybrid voice transcription provides manual review and transcription for at least one utterance from a voice stream and assigning a transcribed value to similar utterances from other voice streams.

Hybrid voice transcription requires a system involving automated voice recognition with human intervention. FIGURE 1 is a functional block diagram showing a system for hybrid voice transcription in accordance with one embodiment. A speech recognition server 14 is interconnected to conventional telephone handsets 12 through Plain Old Telephone Service (POTS) 13, portable handsets 19, through cellular and satellite telephone service 18, and desktop 14 or portable 16 computers, including VoIP clients, Internet clients and Internet telephony clients, through an internetwork 17, such as the Internet. The speech recognition server 14 includes an internetwork 17, such as the Internet. The speech recognition server 14 receives a voice stream 22 from a user via the conventional telephone handsets 12, portable handsets 19, and computers 14, 16. The voice stream can be provided as a real-time voice stream 19 and computers 14, 16. The voice stream can be provided as a real-time voice stream or as a recorded voice message. The speech recognition server 14 includes a parser 24, scorer 25, pool generator 26, and message generator 27. The parser 24 identifies the voice stream 22 received on the speech recognition server 14 and parses the stream into utterances, which can include word-level tokens, n-grams, raw terms, and parse the stream into utterances, which can include word-level tokens, n-grams, raw terms, noun phrases, and sentences. Other types of utterances are possible. The scorer 25 analyzes the utterances and assigns an initial transcribed value and confidence score to each utterance. The initial transcribed values can be selected from a grammar 23 that is stored in a database 20 coupled to the speech recognition server 20. The transcribed values are combined to generate a transcribed message 23 for the voice stream. The pool generator 26 identifies those utterances in the initial transcribed message with a low confidence score and attempts to group those utterances in the initial transcribed message with a low confidence score and attempts to group those...
utterances with similar utterances into a pool. Additionally, the pool generator 26 can select a sample of utterances from the pool and transmit the sample to at least one human transcriber 15 for further processing and analysis. The human transcriber 14 can listen to each utterance in the sample and confirm the initial transcribed value or assign an additional transcribed value. In a further embodiment, the additional processing of the sample is performed automatically, rather than manually. The transcriber can be an employee of a call center, a third party transcription service, or other business.

The manually transcribed value for each utterance in the sample are transmitted to the speech recognition server 11 and compared. If the manually transcribed values differ, the speech recognition server 11 and compared. If the manually transcribed values differ, the remaining similar utterances in the pool are also manually transcribed. However, if all manually transcribed values are the same, the common transcribed value is assigned to the remaining similar utterances in the pool. Subsequently, the message generator 27 incorporates the manually assigned or affirmed transcribed value into the transcribed message for the voice stream 22 from which the utterance was obtained. The transcribed message 23 and corresponding voice stream 22 can be stored in the database 20 for further reference. Further, the transcribed message 23 is provided to the user.

The handsets 12, 19, computers 14, and speech recognition server 11 each include components, conventionally found in general purpose programmable computing devices, such as a central processing unit, memory, input/output ports, network interfaces, and non-volatile storage, although other components are possible. Moreover, other information sources in lieu of or in addition to the servers, and other information consumers, in lieu of or in addition to the handsets and computers, are possible.

Additionally, the handsets 12, 19, computers 14, 16, and speech recognition server 11 can each incorporate one or more modules for carrying out the embodiments disclosed herein. The modules can be implemented as a computer program or procedure written as source code in a conventional programming language and is presented for execution by the central processing unit object or byte code. Alternatively, the modules could also be implemented in hardware, either as object or byte code. Alternatively, the modules could also be implemented in hardware, either as integrated circuitry or burned into read-only memory components. The various implementations of the source code and object and byte codes can be held on a computer-readable storage medium, such as a floppy disk, hard drive, digital video disk (DVD), random access memory (RAM), read-only memory (ROM) and similar storage mediums. Other types of modules and module functions are possible, as well as other physical hardware components of module The hybrid voice transcription system 10 can be incorporated for use in other systems, including a call center. FIGURE 2 is a functional block diagram showing, by way of example, including a call center. FIGURE 2 is a functional block diagram showing, by way of example,
the hybrid voice transcription system of FIGURE 1 as incorporated into a call center 30. The hybrid voice transcription system of FIGURE 1 has been incorporated into a call center 30. The call center 30 includes a message server 31 and a speech recognition server 34, operatively interconnected over a network 33, such as an internal network or the Internet. One or more human transcribers 32a-c can also be interconnected to the message server 31 and speech recognition server 34 over the network 33. The human transcribers 32a-c can be employees of the call center 30, as well as employees of a third party service provider. In a further embodiment, the human transcribers 32a-c can be replaced with an automated system for further processing the transcribed messages, including assigning further transcription values to one or more utterances. The network infrastructure can be either wired or wireless and, in one more utterances. The network infrastructure can be either wired or wireless and, in one embodiment, is implemented based on the Transmission Control Protocol/Internet Protocol (TCP/IP) network specification, although other types or combinations of networking arrangements are possible. Similarly, other network topologies and arrangements are possible.

Incoming customer calls are received through a call interface 35, which is operatively coupled to the message server 31 to provide access to a telephone voice and data network. In one embodiment, the call center interface 35 connects to the telephone network over a T1-carrier line, which can provide up to 24 individual channels of voice or data traffic provided at 64 kilobits (Kbits) per second. Other types of telephone network connections are possible. Each received message server 31 sends streamed audio data from the customer call as a voice stream 36 to the speech recognition server 34, which performs automatic speech recognition by parsing the voice stream into utterances and assigning a transcribed value and confidence score to each utterance.

An accuracy threshold is applied to the confidence scores of the utterances and those utterances with confidence scores below the threshold are selected as questionable utterances. Utterances with confidence scores below the threshold are selected as questionable utterances. Questionable utterances can include those utterances for which the initial transcribed value has a threshold likelihood of being an incorrect representation of that utterance. Higher confidence scores indicate a higher likelihood that an initial transcribed value assigned to an utterance is an accurate transcription of that utterance. Whereas, lower confidence scores indicate a lower likelihood that an initial transcribed value is an accurate transcription of an utterance. The value of the confidence score assigned reflects the degree of likelihood that an initial transcribed value reflects the corresponding utterance. One or more of the questionable utterances can be transmitted for further analysis, such as by a human transcriber, for confirming the initial transcribed value or for assigning a further transcribed value. Hybrid voice transcription will be transcribed value or for assigning a further transcribed value. Hybrid voice transcription will be
discussed below with respect to a call center. However, hybrid voice transcription can also be discussed below with respect to a call center. However, hybrid voice transcription can also be incorporated for use in other systems for fields, such as the legal or medical fields. As briefly described above, hybrid voice transcription can occur through a sequence of phases. FIGURE 3 is a process flow diagram showing a method 40 for hybrid voice transcription, in accordance with one embodiment. Audio data is received from a caller as a voice stream, which is parsed into utterances (block 41). Each utterance includes tokens, such as voice stream, which is parsed into utterances (block 41). Each utterance includes tokens, such as words, n-grams, raw terms, noun phrases, and sentences. Other types of utterances are possible. A transcribed value and confidence score are initially assigned to each utterance via automated speech recognition (block 42). The initial transcribed value can include a text representation of the corresponding voice utterance and can be automatically selected from a grammar, which includes a list of common transcribed values that may occur in the voice stream during a call. One or more grammars can be selected based on characteristics of a caller, including geographic location, age, interests, and network. Additionally, the confidence score provides a measure of certainty that the transcribed value accurately represents the voiced utterance. An accuracy threshold is applied to the confidence scores of the utterances and those utterances with confidence scores below the threshold are selected for further analysis (block 43). The threshold can be automatically determined or manually set, such as by an employee of a call center, including a manager. At least one of the questionable utterances is selected and attempts to build a pool of similar questionable utterances for the selected questionable utterance are made (block 44). If a pool is formed, one or more of the similar questionable utterances can be selected as a sample for further analysis. The information from the further analysis can be used to confirm the accuracy of or change the initial transcribed value assigned to the remaining similar questionable utterances in the pool without having to be individually reviewed by a manual transcriber. Building the pool of similar questionable utterances can include identifying a predetermined number of similar questionable utterances within a predetermined time frame. Identifying similar questionable utterances for inclusion in a pool is further discussed below with reference to FIGURE 4. If an appropriate sized pool of similar questionable utterances are not retrieved to FIGURE 4. If an appropriate sized pool of similar questionable utterances are not retrieved to FIGURE 4. If an appropriate sized pool of similar questionable utterances are not retrieved to FIGURE 4. If an appropriate sized pool of similar questionable utterances are not retrieved to FIGURE 4. If an appropriate sized pool of similar questionable utterances are not retrieved to FIGURE 4. If an appropriate sized pool of similar questionable utterances are not retrieved to FIGURE 4. If an appropriate sized pool of similar questionable utterances are not retrieved to FIGURE 4. If an appropriate sized pool of similar questionable utterances are not retrieved to FIGURE 4. If an appropriate sized pool of similar questionable utterances are not retrieved to FIGURE 4. If an appropriate sized pool of similar questionable utterances are not retrieved to FIGURE 4. If an appropriate sized pool of similar questionable utterances are not retrieved to FIGURE 4. If an appropriate sized pool of similar questionable utterances are not retrieved to FIGURE 4. If an appropriate sized pool of similar questionable utterances are not retrieved to FIGURE 4. If an appropriate sized pool of similar questionable utterances are not retrieved to FIGURE 4. If an appropriate sized pool of similar questionable utterances are not retrieved to FIGURE 4. If an appropriate sized pool of similar questionable utterances are not retrieved to FIGURE 4. If an appropriate sized pool of similar questionable utterances are not retrieved to FIGURE 4. If an appropriate sized pool of similar questionable utterances are not retrieved to FIGURE 4. If an appropriate sized pool of similar questionable utterances are not retrieved to FIGURE 4. If an appropriate sized pool of similar questionable utterances are not retrieved to FIGURE 4. If an appropriate sized pool of similar questionable utterances are not retrieved to FIGURE 4. If an appropriate sized pool of similar questionable utterances are not retrieved to FIGURE 4. If an appropriate sized pool of similar questionable utterances are not retrieved to FIGURE 4.
Once selected, the sample of similar questionable utterances is transmitted to a human transcriber for confirming the assigned transcribed values or assigning further transcribed values to each utterance in the sample. The questionable utterances in the sample can be transmitted to the same human transcriber or, alternatively, to different human transcribers. Each similar questionable utterance in the sample is analyzed to identify the correct transcribed value associated with that utterance. Upon receipt from the human transcriber, the confirmed or newly assigned transcribed values are compared. If all questionable utterances in the sample are assigned transcribed values are compared. If all questionable utterances in the sample are associated with the same transcribed value (block 49), the common transcribed value is then assigned to the questionable utterances in the pool from which the confirmed or assigned (block 47) to all the questionable utterances in the pool which the sample was selected. The common transcribed value is confirmed for those questionable samples selected. The common transcribed value is confirmed for those questionable utterances having initial transcribed values that match the common transcribed value. The common transcribed value is assigned to the questionable utterances when the initial transcribed value is different from the common transcribed value. If no common transcribed value is provided for the similar questionable utterances in the sample, each remaining questionable utterance in the pool is then transmitted to the human transcriber (block 48) for further review, such as confirmation of the initial transcribed value or assignment of a further transcribed value.

The transcribed values confirmed or assigned by the human transcriber can be incorporated into the initial transcribed message (block 49) for the corresponding utterance, which can be provided to the caller (block 50). In a further embodiment, the transcribed value and associated utterance can be stored with the grammar or used to generate a further grammar.

During hybrid voice transcription, a subset of similar questionable utterances can be identified and transmitted for further processing to increase the accuracy of fully automated identification and transmitted for further processing to increase the accuracy of fully automated speech recognition systems, while maintaining efficiency and reducing cost. The subset of similar questionable utterances can be generated as a pool from which a sample can be selected. Similar questionable utterances can be generated as a pool from which a sample can be selected. FIGURE 4 is a process flow diagram showing, by way of example, a method 60 for building the pool of similar questionable utterances for use in the process of FIGURE 4. Questionable utterances are identified (block 61) from an initial transcribed message based on a confidence score for a corresponding initial transcribed value assigned to each utterance. In one embodiment, the questionable utterances include those utterances with initial transcribed values having confidence scores that fall below a predetermined accuracy threshold. At least one of the questionable utterances is selected and monitored for a predetermined amount of time during which questionable utterances in voice streams from other calls are watched. The other
questionable utterances that are similar to the selected questionable utterance are identified (block 62) and grouped with the questionable utterance (block 63). The similarity of two or more questionable utterances can be determined based on similarity metrics. FIGURE 5 is a data flow diagram showing, by way of example, similarity metrics 7.1 and 7.2 for use in the process of FIGURE 4. The similarity metrics 7.1 include factors for determining whether two or more questionable utterances are similar. Similarity metrics 7.1 can be applied to require that a grouping of similar questionable utterances reaches a certain size to be monitored. The similarity of two or more questionable utterances can be determined, based on similarity metrics 7.1, for use in the process of FIGURE 4. The similarity metrics 7.1 include factors for determining whether two or more questionable utterances are similar, such as by the initial transcribed values. Confidence scores 7.3, range of similarity 7.4, shared between the at least two transcribed values, and call characteristics 7.5. For example, using the initial transcribed two transcribed values, and call characteristics 7.5. For example, using the initial transcribed value similarity factor, if two utterances have each been assigned the same initial transcribed value similarity factor, if two utterances have each been assigned the same initial transcribed values, the two utterances can be determined to be similar. Alternatively, even if the initial values, the two utterances can be determined to be similar. Alternatively, even if the initial transcribed values are different, they may still be considered to be similar if they fall within a similarity range of transcribed values. The identification of similar questionable utterances can also be based on the confidence score associated with the initial assigned transcribed values. For example, if two utterances are assigned a common initial transcribed value, the two utterances may only be considered similar if the associated confidence scores fall above a confidence score threshold or within a particular range of confidence.

Also, a range of similarity 7.4 can also be applied to determine whether the initial transcribed value or the corresponding confidence score for a questionable utterance are within the predetermined bounds for determining similarity with another questionable utterance. As well, call characteristics, such as geographic area, age, and affiliations can be used as factors in determining whether two questionable utterances are similarly related. Further examples of determining similarity based on the similarity metrics are provided below with reference to FIGURE 8.

During the identification of similar questionable utterances, at least two factors are monitored to determine whether the grouped similar questionable utterances form a pool from which a sample can be selected for further analysis. The factors include a time threshold (block 64) and a size threshold (block 65). Regarding the time limit threshold, an automatically determined or predetermined time period is selected during which a pool of similar questionable utterances must be grouped. If grouped within the selected time period (block 64), the similar utterances are recognized as a pool from which a sample of the questionable utterances can be selected. However, if not grouped within the time period, each similar questionable utterance identified is transmitted to the human transcriber for individual analysis. The time period can be an absolute time or a period of time that covers a range. Additionally, the size threshold can be applied to require that a grouping of similar questionable utterances reaches a certain size to be applied to require that a grouping of similar questionable utterances reaches a certain size to be
The size threshold can be automatically determined or predetermined to establish a particular number of similar questionable utterances that must be grouped to form a pool, from which a sample can be selected. If the size threshold is satisfied (block 65), one or more of the similar questionable utterances can be selected as a sample for further analysis, whereas each of the similar questionable utterances is transmitted for further analysis when the size threshold is not satisfied.

Setting the size and time thresholds can be determined based on a type of voice stream, accuracy of transcription required, and type of service agreement signed. The type of voice accuracy of transcription required, and type of service agreement signed. The type of voice stream can include a real-time voice stream or a recorded voice message. When the utterances stream can include a real-time voice stream or a recorded voice message. When the utterances are being provided in a real-time voice stream during a live call, the time to identify similar are being provided in a real-time voice stream during a live call, the time to identify similar questionable utterances should be shorter, whereas when the caller is waiting on the call. In one embodiment, the time limit can be set at three to thirty seconds for a real-time call. However, when the utterances are provided in a recorded voice message, the time to identify similar questionable utterances can be longer. In one embodiment, the time limit can be set at 30 minutes for a recorded voice message. With respect to live calls, the size threshold can be set lower, such that fewer similar utterances are required to form a pool. Additionally, for recorded calls, the size threshold can be set higher since more time is provided to identify similar questionable utterances.

Further, the size threshold for a pool can be determined based on an accuracy of the transcription required. For example, a larger pool provides more accurate results, whereas a smaller pool provides more relaxed results. High accuracy transcription may be required in court reporting and lower accuracy may be required for simple consumer calls that request information.

In one embodiment, the time limit is monitored and upon expiration, the identified similar utterances are analyzed to determine whether the size threshold is satisfied. If the pool fails to satisfy either the size threshold or the time limit, no pool of the similar questionable utterances is formed. However, if a pool of similar questionable utterances satisfies the size threshold and is formed within the predetermined time limit, the pool is assigned as a pool for use in the hybrid voice transcription.

Once a pool of similar questionable utterances is generated, a sample can be selected for sending to the human transcriber. FIGURE 6 shows a data flow diagram 60 showing, by way of example, metrics for obtaining a sample 81 from the pool of FIGURE 40. A sample can be obtained from a pool based on random sampling 82, specific sampling 83, or a combination of random and specific sampling 84. In a random sampling, one or more questionable

- 10 -
utterances are selected from the pool at random. The number of selected questionable utterances can be automatically determined or determined by a user. During specific sampling, particular questionable utterances are selected as representative of the sample based on factors, including confidence score, characteristics of the caller, subject matter, and noise, as well as other factors. A combination of random sampling and specific sampling can also be used. For instance, a predetermined number of questionable utterances can be selected through random sampling, a further set of questionable utterances can be selected through specific sampling to generate a complete sample. In a further embodiment, random sampling can be used to identify a sample of questionable utterances, which can be further refined through specific sampling to generate a sample of questionable utterances, which can be further refined through specific sampling. 

Other combinations of random sampling and specific sampling are possible. Other combinations of random sampling and specific sampling are possible. This allows for the size of a sample to be determined based on an accuracy of transcriptions. The size of a sample can be determined based on an accuracy of the transcription required. For example, a larger sample provides more accurate results, whereas a smaller sample provides more relaxed results. High accuracy transcription may be required in legal proceedings, such as court reporting, while lower accuracy results may be sufficient for simple consumer calls. 

FIGURE 7 illustrates a block diagram showing, by way of example, a transcribed message. The transcribed message includes a message details box and a text box. The message details box includes fields for message characteristics, including, for example, message identification, date, time, sender, and recipient. Other message characteristics are possible, such as identification of related messages or time required to transcribe the message. 

The message identification field can include a number, letter, or symbol, for representing a transcribed message for a particular voice stream. Corresponding voice streams and transcribed messages can have the same or different identification values, as well as related identification values, as well as related identification values. The date field can include the date on which the original voice stream was received, the date on which the voice stream was transcribed, or the date on which the received, the date on which the voice stream was transcribed, or the date on which the transcribed message was provided to a caller. In one embodiment, the date for receiving the transcribed message was provided to a caller. In one embodiment, the date for receiving the voice stream, for transcribing the voice stream, and for providing the transcribed message is the same. Similar to the date field, the time field can include the time at which the original voice stream was received, the time at which the voice stream was transcribed, or the time at which the transcribed message was provided to a caller. 

The present invention can also include the name of an individual, business entity, or other recipient. Additionally, other identifier(s) can be used, such as an identification number for a sender, telephone number of the sender, etc. Similarly, the recipient field can include the name of an individual, a business entity, or other recipient, as well as other identifiers, such as an individual, a business entity, or other recipient, as well as other identifiers, such as an individual, a business entity, or other recipient, as well as other identifiers, such as an individual, a business entity, or other recipient, as well as other identifiers, such as an
identification number or telephone number. Other identifiers are possible. As well, the recipient field can include one or more recipients, each identified by the same or different types of identifiers. In the text box 92, an initial message 94 transcribed via automated speech recognition can be provided. In the transcribed message 94, each utterance in the corresponding voice stream is assigned a transcribed value, which is represented as text. The transcribed message 94 in the text box reads:

Good Morning "hector" - it's Jeff - just about "eigtam" to follow
up with you on the - I "ATT" awards nomination I don't have "qcall" back from -...", I think she'd asked you to get involved - I've "gaaat"?
get that done today

Each transcribed utterance can be associated with a confidence score (not shown), which reflects a level of certainty as to whether the transcribed value accurately reflects the voiced utterance.

An accuracy threshold can be applied to the confidence scores of each utterance in the initial transcribed message 94 to identify those utterances with transcribed values having confidence scores below the threshold as questionable utterances. If above the threshold, the transcribed values assigned to the utterances can be considered accurate or fairly accurate. Meanwhile, the questionable utterances can have a likelihood of being associated with an incorrect transcribed value. The confidence score reflects the likelihood of accurate transcribed values. In the transcribed message 94, the questionable utterances can be identified by quotation marks. In place of, or in addition to, the quotation marks, the questionable utterances can also be displayed by highlighting one or more utterances, displaying one or more utterances can also be displayed by using one or more utterances, displaying one or more utterances in an utterance box, and through different font sizes and styles.

The transcribed values for the questionable utterances can each be represented by identification values for use in identifying similar questionable utterances. FIGURE 8 is a block diagram showing, by way of example, a transcribed message with identification values representing transcribed values for questionable utterances. Each transcribed value associated with a questionable utterance in the transcribed message is replaced with an identification value and a confidence score for the transcribed value. The identification value can include a number, code, or symbol. Each transcribed value is assigned a different identification value; however, the same or similar transcribed values may be assigned the same identification value. Other types of and assignments of identification values are possible. The questionable utterances can be displayed with quotation marks or other displays, including highlighting, font size, and utterance displayed with quotation marks or other displays, including highlighting, font size, and utterance.
boxes. For instance, in the transcribed message, the transcribed values having a confidence score above the threshold can be represented as text, which is highlighted in a first color, and the questionable utterances can be represented as text that is highlighted in a second color. Accordingly, the two utterances are considered similar.

One or more of the questionable utterances can be selected and watched for a predetermined amount of time during which attempts are made to locate similar questionable utterances. Returning to the above example, a threshold of 80% is applied to the transcribed values in the message. Those utterances associated with transcribed values having a confidence level lower than the threshold are selected as questionable utterances. Thus, the utterances level lower than the threshold are selected as questionable utterances. Those utterances associated with transcribed values identified by the values "111100," "X2000," "Y4359," "X45600," "332200," "222200," and "X3450," are selected as questionable utterances. Further, the questionable utterance associated with the transcribed value "gaaata," is selected. "Gaaata" is the questionable utterance associated with the transcribed value "gaaata" is selected. "Gaaata" is represented by the identification value "X3450" with a confidence score of 65%. The corresponding questionable utterance is then watched for a time period of 30 minutes during which attempts to locate similar questionable utterances are made.

The located similar questionable utterances form a pool with the selected questionable utterance.

As described above with reference to FIGURE 4, the pool can be generated from a predetermined number of similar questionable utterances, which are identified within a predetermined amount of time. Each questionable utterance within the pool is represented by an identification value, which is based on a transcribed value, and a confidence score. The similarity between questionable utterances can be determined based on similarity metrics, including an assigned transcribed utterance, confidence score, range of similarity metrics, including an assigned transcribed utterance, confidence score, range of similarity, and call characteristics. However, other similarity metrics are possible. Returning to similarity, and call characteristics, however, other similarity metrics are possible. Returning to the above example, the transcribed value "gaaata" is watched while a search is performed for the above example, the transcribed value "gaaata" is watched while a search is performed for the two questionable utterances. A further questionable utterance with the transcribed similar questionable transcribed utterances. A further questionable utterance with the transcribed value "gotta" is identified. The further transcribed value is assigned an identification value of value "gorta" is identified. The further transcribed value is assigned an identification value of "X435 1" with a confidence score of 55%. A similarity comparison of the transcribed values is performed between the two questionable utterances based on one or more of the similarity metrics. For instance, in this example, the applicable similarity metrics include the assigned transcribed value, confidence score, and range of similarity. The transcribed value "gaaata" is associated with the identification value of "X3450," while "gotta" has an identification value of "X345 1." The assigned transcribed value similarity metric requires that the identification value of two or more utterances fall within a range of five. In this case, the identification values of the two utterances differ by a value of one. Accordingly, the two utterances are considered similar.
For the confidence score comparison, an automatic threshold or a predetermined threshold can be used to identify similar questionable utterances. For instance, a confidence score threshold of 35% is selected and those utterances with similar identification values, as described above, that have confidence scores over 35% can be selected as similar. Alternatively, use of the bounded range similarity metric requires that the identification value, confidence score, or both the identification value and confidence scores fall within a predetermined bounded range, both the identification value, and confidence scores fall within a predetermined bounded range, including an upper and lower limit for two or more utterances to be considered similar.

Also, call characteristics can be used as factors for determining similarity between two or more questionable utterances, including any aspect of a call or the caller, such as location, topic, more questionable utterances, including any aspect of a call or the caller, such as location, topic, emotion, accent, gender, and age. Other call characteristic factors are possible. The call emotion, accent, gender, and age. Other call characteristic factors are possible. The call characteristics can be provided by a caller or obtained from an analysis of the call, such as by caller ID.

In one example, an utterance with a transcribed value of "U-dub" is parsed from a voice stream. A caller in Oregon generated the voice stream during a call to a call center. The location of the caller can be identified via caller ID or directly provided by the caller. Alone, the transcribed value "U-dub" could refer to the University of Washington or the University of Wisconsin. Consideration of location information can help distinguish the correct reference for "U-dub." A caller located in the Pacific Northwest, such as Washington or Oregon, is more likely to refer to the University of Washington. However, a caller in the Midwest, such as Wisconsin, is more likely to use "U-dub" as a reference to the University of Wisconsin. Accordingly, in the current example, "U-dub," will have a higher similarity to other references to the University of Washington, rather than references to the University of Wisconsin. Additionally, utterances with a transcribed value of "U-dub" from a caller in Oregon and caller in Wisconsin would not be considered similar. In Wisconsin would not be considered similar. Additionally, in a further example, age can be used as a factor for grouping similar utterances that represent slang. "Scrilla" is a term of slang that is used by youth and young adults to refer to money. Thus, knowledge of a caller's age can help determine whether the transcription of the utterance is complete or inaccurate. Accordingly, the two transcribed utterances would not be considered similar. A pool of similar questionable utterances has been formed, a sample of the questionable utterances can be selected for transmitting to one or more human transcribers, as questionable utterances can be selected for transmitting to one or more human transcribers, as
described above in detail with reference to FIGURE 6. Each questionable utterance in the
sample is individually transcribed via manual transcription. If the same transcribed value is
assigned to each questionable utterance in the sample, the common transcribed value is assigned
to each of the remaining questionable utterances in the pool. When the common transcribed
value is different than the initial transcribed value, the assigned common value replaces the
initially assigned transcribed value. The transcribed message to which the questionable utterance
belongs is also updated to reflect the common transcribed value. However, if the common
transcribed value is the same as the initially transcribed value, the initially transcribed value is
confirmed as being a correct representation of the associated utterance. If different transcribed
values are assigned to the questionable utterances in the sample, each of the remaining
questionable utterances in the sample are manually transcribed.

The manually assigned transcribed values can be stored for further reference. FIGURE 10 is a block diagram showing, by way of example, a transcribed value chart 120. The
transcribed value chart 120 includes questionable utterances identified for at least one voice
stream, which are each represented by an identification value and confidence score 123. The
identification value and confidence score can be assigned during automated speech recognition.
Additionally, the transcribed value chart 120 includes the actual transcribed values 122, which
include the transcribed values 124 provided by the human transcriber or the transcribed value
provided by automated speech recognition. The actual transcribed values have
been verified as having a high likelihood of correctly representing the corresponding utterance.
In one embodiment, the questionable utterances and transcribed values can be stored separately
for each voice stream or alternatively, the questionable utterances and transcribed values can be
stored for more than one voice stream.

In a further embodiment, the voice utterance and assigned transcribed value can be added
to the grammar, which is used during automated speech recognition to assign initial transcribed
to the grammar, which is used during automated speech recognition to assign initial transcribed
values to the utterances in the voice stream, as described above in FIGURE 3. In yet a further
embodiment, the voice utterance and assigned transcribed value are only added to the grammar
embodiment after the transcribed value has been assigned to the utterance on a predetermined number of
occasions. For instance, a predetermined number of occurrences can be set at two. Returning to
the above example, an utterance is initially assigned a transcribed value of "gaaata." However,
after additional analysis, the utterance is assigned a further transcribed value of "gotta." A further utterance is subsequently received, which is the same or similar to the previously received utterance. The further utterance is also assigned an initial transcribed value of "gaaata." Upon further analysis, the initial transcribed value is replaced with a subsequent transcribed value.
value “gotta.” Since the same or similar utterances have been twice assigned the transcribed value “gotta,” the new transcribed value can then be added to the grammar. Once added, the next time the same or similar utterance is received, the term “gotta” is initially assigned as the transcribed value without having to undergo further analysis, such as by a human transcriber.

In yet a further embodiment, the voice utterance and assigned transcribed value can be added to a secondary grammar. The addition of a transcribed value to a secondary grammar can be effective when the transcribed value is not commonly used enough to warrant addition to the main grammar or when the addition of the transcribed value would deteriorate the effectiveness of the main grammar or when the addition of the transcribed value would deteriorate the effectiveness of a similar transcribed value on the main grammar that is more important. Returning to the above example, instead of being added to the grammar, the transcribed value “gotta” is assigned to a secondary grammar since the term “gotta” is slang for the phrase “got to.” The next time the same or similar utterance is received, the utterance is compared with the main grammar and then the secondary grammar from which the term “gotta” is initially assigned as the transcribed value without having to undergo further analysis, such as by a human transcriber. As described above, the assignment of the transcribed value may be required to occur a particular number of times before the transcribed value is added to the secondary grammar. Additional grammars are also possible, including third, fourth, and fifth grammars.

While the invention has been particularly shown and described as referenced to the embodiments thereof, those skilled in the art will understand that the foregoing and other changes in form and detail may be made therein without departing from the spirit and scope of the invention.
CLAIMS:
1. A system (10) for efficiently reducing transcription error via hybrid voice transcription, comprising:
a parser (24) to parse a voice stream (22) from a call into utterances (41) and assigning an initial transcribed value (42) and corresponding recognition score (113) to each utterance;
a message generator (27) to generate a transcribed message (23) for the call comprising the initial transcribed values;
a threshold module to apply a threshold to the recognition scores (113) to identify those utterances with recognition scores (113) below the threshold as questionable utterances (111); and
a comparison module to compare at least one questionable utterance (111) to other questionable utterances (111) from other calls and forming a group of similar questionable utterances (111); and
an assignment module to select one or more of the similar questionable utterances (111) from the group, to receive a common manual transcription value (124) for the selected similar questionable utterances (111), and to assign the common manual transcription value (124) to the remaining similar questionable utterances (111) in the group.

2. A system (10) according to Claim 1, further comprising:
a replacement module to replace the initial transcribed value for the at least one questionable utterance with the common manual transcription value (124) in the transcribed message (23).

3. A system (10) according to Claim 1, further comprising:
a pool generator to build the group of similar questionable utterances (111) based on at least one of a size threshold and a time threshold.

4. A system (10) according to Claim 1, further comprising:
a sample module to select a sample (81) comprising the one or more similar questionable utterances (111) based on one of random sampling (82), specific sampling (83), and a combination (84) of random and specific sampling, and a combination (84) of random and specific sampling.
5. A system (10) according to Claim 1, further comprising:

- A similarity module (11) to determine a similarity (71) between the at least one questionable utterance (111) and the other questionable utterances (111) based on at least one of the initial transcribed value (72), recognition score (113), range of similarity (74), and characteristics of the call (75).

6. A method (40) for efficiently reducing transcription error via hybrid voice transcription, comprising:

- Parsing a voice stream (22) from a call into utterances (41) and assigning an initial transcribed value and corresponding recognition score (113) to each utterance;
- Generating a transcribed message (23) for the call comprising the initial transcribed values;
- Applying a threshold to the recognition scores (113) to identify those utterances with recognition scores (113) below the threshold as questionable utterances (111);
- Comparing at least one questionable utterance (111) to other questionable utterances (111) from other calls and forming a group of similar questionable utterances (111);
- Selecting one or more of the similar questionable utterances (111) from the group and receiving a common manual transcription value (124) for the selected similar questionable utterances (111); and
- Assigning the common manual transcription value (124) to the remaining similar questionable utterances (111) in the group.

7. A method (40) according to Claim 6, further comprising:

- Replacing the initial transcribed value for the at least one questionable utterance with the common manual transcription value (124) in the transcribed utterance in the transcribed message (23).

8. A method (40) according to Claim 6, further comprising:

- Building the group of similar questionable utterances (111) based on at least one of a size threshold and a time threshold.

9. A method (40) according to Claim 6, further comprising:
selecting a sample (81) comprising the one or more similar questionable utterances (111) based on one of random sampling (82), specific sampling (83), and a combination (84) of random and specific sampling.

10. A method (40) according to Claim 6, further comprising:
   determining a similarity (71) between the at least one questionable utterance and the other questionable utterances (111) based on at least one of the initial transcribed value (72), recognition score (113), range of similarity (74), and characteristics of the call (75).

11. A system (10) for hybrid voice transcription error reduction, comprising:
   a parser (24) to parse a call into speech utterances (41);
   a message generator (27) to generate a transcribed message (23) for the call by assigning an initial transcribed value and corresponding recognition score (113) to each utterance;
   a threshold module to apply a confidence threshold (113) to the recognition scores (113) and to select those utterances with initial transcribed values that fall below the threshold as questionable utterances (111);
   a sample module to generate a sample (81) for at least one of the questionable utterances (111) comprising identifying other questionable utterances (111) from other calls that are similar to the at least one questionable utterance (111), grouping the at least one questionable utterance (111) and similar questionable utterances (111) as related utterances, and selecting one or more of the related utterances as the sample (81); and
   a module to receive a common manual transcription value (124) for each of the related utterances in the sample (81) and assigning the common manual transcription value (124) to the remaining related utterances in the group.

12. A system (10) according to Claim 11, further comprising:
   a pool generator (26) to form the group of the related utterances by identifying a threshold number of the related utterances within a predetermined time limit, number of the related utterances within a predetermined time limit.
13. A system (10) according to Claim 11, further comprising:

- a replacement module to replace the initial transcribed value with the common manual transcription value (124) for the at least one questionable utterance in the transcribed message (23).

14. A system (10) according to Claim 11, wherein the sample (81) is selected using at least one of random sampling (82), specific selection sampling (73), and a combination (84) of random and specific selection sampling.

15. A system (10) according to Claim 11, further comprising:

- a similarity module to determine the similarity (71) of the at least one questionable utterance (111) and the other questionable utterances (1111) based on one or more of the initial transcribed value (72), recognition score (113), range of similarity between the related utterances (74), and characteristics of the call (75).

16. A method (40) for hybrid voice transcription error reduction, comprising:

- parsing a call into speech utterances (41) and generating a transcribed message (23) for the call by assigning an initial transcribed value and corresponding recognition score (113) to each utterance;
- applying a confidence threshold to the recognition scores (113) and selecting those utterances with initial transcribed values that fall below the threshold as questionable utterances (111);
- generating a sample (81) for at least one of the questionable utterances generating a sample (81) for at least one of the questionable utterances (1111), comprising:
  (1111), comprising:
  - identifying other questionable utterances (111) from other calls that are similar to the at least one questionable utterance (111);
  - identifying other questionable utterances (111) from other calls that are similar to the at least one questionable utterance (111);
  - grouping the at least one questionable utterance (111) and similar questionable utterances (111) as related utterances;
- selecting one or more of the related utterances as the sample (81);
receiving a common manual transcription value (124) for each of the
related utterances in the sample (81) and assigning the common manual
transcription value (124) to the remaining related utterances in the group.

17. A method (40) according to Claim 16, further comprising:
forming the group of the related utterances by identifying a threshold
number of the related utterances within a predetermined time limit.

18. A method (40) according to Claim 16, further comprising:
replacing the initial transcribed value (124) with the common manual
transcription value for the at least one questionable utterance (111) in the
transcribed message (23).

19. A method (40) according to Claim 16, wherein the sample (81)
is selected using at least one of random sampling (82), specific selection
sampling (83), and a combination (84) of random and specific selection
sampling.

20. A method (40) according to Claim 16, further comprising:
determining the similarity (71) of the at least one questionable
utterance (111) and the other questionable utterances (111) based on one or
more of the initial transcribed value (124), recognition score (113), range of
similarity between the related utterances (74), and characteristics of the call
(75).
Fig. 3.

Start

1. Parse voice stream into utterances
2. Assign a transcribed value and confidence score to each utterance
3. Apply a threshold to the confidence scores and select those utterances of confidence scores below the threshold

4. Build pool for at least one questionable utterance?
   - Yes: Select sample from pool
   - No: Proceed to next step

5. Receive same transcribed value for each utterance?
   - Yes: Assign transcribed utterance to all questionable utterances in pool
     - Incorporate assigned transcribed value to transcribed message
     - Provide transcribed message to caller
   - No: Proceed to next step

6. Receive transcribed value for each identified utterance in pool

End
Fig. 4.

60 Build Pool

61 Identify questionable utterance

62 Identify similar questionable utterances from other voice messages

63 Group similar questionable utterances with identified questionable utterance

64 Group formed within time limit?
   Yes
   65 Group size satisfy size threshold?
      Yes
      66 Assign group as pool
      End
      No
      67 No pool is formed
      End
   No
   67 No pool is formed
   End
Fig. 5.

- Call characteristics
- Assigned transcribed value
- Similarity metrics
- Range of similarity
- Confidence score

Fig. 6.

- Random
- Sample
- Combination
- Specific
Good Morning “hector” - it’s Jeff - just about “eigtam” to follow up with you on -the -I “ATT” awards nomination – I don’t have “qcall” back from - “” I think she’d asked you to get involved - I’ve “gaaata”? get that done today
Good Morning "111100-75%" - it’s Jeff
- just about “X2000-50%”-
“Y4359-65% to follow up with you on
-the -I “X45600-75%” awards
Nomination – I don’t have
“333200-57%” back from -
“222200-65%” I think she’d asked
You to get involved - I’ve
“X3450-65%”? Get that done today
<table>
<thead>
<tr>
<th>Machine Variable</th>
<th>Actual Word</th>
</tr>
</thead>
<tbody>
<tr>
<td>111100 - 75%</td>
<td>Howard</td>
</tr>
<tr>
<td>X2000 - 50%</td>
<td>8:00 am</td>
</tr>
<tr>
<td>Y4359 - 65%</td>
<td>Hey</td>
</tr>
<tr>
<td>X45600 - 75%</td>
<td>ATT</td>
</tr>
<tr>
<td>333200 - 57%</td>
<td>Anything</td>
</tr>
<tr>
<td>222200 - 65%</td>
<td>Heidi</td>
</tr>
<tr>
<td>X3450 - 65%</td>
<td>Gotta</td>
</tr>
</tbody>
</table>
**INTERNATIONAL SEARCH REPORT**

**PCT/US2011/043889**

A. CLASSIFICATION OF SUBJECT MATTER

INV. G10L15/06

ADD.

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

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

G10L

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

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

EPO-Internal, INSPEC, WPI Data

C. DOCUMENTS CONSIDERED TO BE RELEVANT

<table>
<thead>
<tr>
<th>Category</th>
<th>Citation of document, with indication, where appropriate, of the relevant passages</th>
<th>Relevant to claim No.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Y</td>
<td>US 2006/212293 AI (TUR GOKHAN [US]) 21 September 2006 (2006-09-21)</td>
<td>1, 2, 4, 6, 7, 9, 11, 13, 14, 16, 18, 19</td>
</tr>
<tr>
<td>Y</td>
<td>RABINER LAWRENCE ET AL: &quot;Cl usteri ng&quot;, FUNDAMENTALS OF SPEECH RECOGNITION, XX, XX, 1 January 1993 (1993-01-01) , pages 267-274, XP002953836, 5.3.3 Cl usteri ng, Fi g. 5.10</td>
<td>1, 2, 4, 6, 7, 9, 11, 13, 14, 16, 18, 19</td>
</tr>
</tbody>
</table>

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 document but published on or after the international filing date
  * "L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
  * "O" document referring to an oral disclosure, use, exhibition or other means
  * "P" document published prior to the international filing date but later than the priority date claimed
  * "T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
  * "X" document of particular relevance; the claimed invention cannot be considered to be obvious when the mentioned document is considered in conjunction with this document
  * "Y" document of particular relevance; the claimed invention cannot be considered to be obvious when the mentioned document is considered in conjunction with another document such document, etc., the combination being obvious to a person skilled in the art
  * "F" document member of the same patent family

**Date of the actual completion of the international search**

28 September 2011

**Date of mailing of the international search report**

11/11/2011

Name and mailing address of the ISA:

European Patent Office, P.B. 5818 Patentlaan 2

NL - 2280 HV Rijswijk

Tel. (+31-70) 340-2040,

Fax: (+31-70) 340-3016

Authorized officer

Dobl er, Ervi n
<table>
<thead>
<tr>
<th>Patent document cited in search report</th>
<th>Publication date</th>
<th>Patent family member(s)</th>
<th>Publication date</th>
</tr>
</thead>
<tbody>
<tr>
<td>US 2006212293 Al</td>
<td>21-09-2006</td>
<td>CA 2540557 Al</td>
<td>21-09-2006</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GB 2424502 A</td>
<td>27-09-2006</td>
</tr>
</tbody>
</table>