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

(11) EP 1 103 954 A1

# (12) EUROPEAN PATENT APPLICATION

(43) Date of publication:
**30.05.2001 Bulletin 2001/22**

(51) Int Cl.[7]: **G10L 19/00**

(21) Application number: **99203979.2**

(22) Date of filing: **26.11.1999**

(84) Designated Contracting States:
**AT BE CH CY DE DK ES FI FR GB GR IE IT LI LU MC NL PT SE**
Designated Extension States:
**AL LT LV MK RO SI**

(71) Applicant: **Lancaster Equities Limited**
**Bridgetown (BS)**

(72) Inventors:
- **Lutz, Marc**
  **North York, Ontario M2N 6C1 (CA)**
- **Vange, Mark**
  **North York, Ontario M2N 6C1 (CA)**

(74) Representative: **Hopfgarten, Nils et al**
**L.A. Groth & Co. KB**
**P.O. Box 6107**
**102 32 Stockholm (SE)**

## (54) Digital speech acquisition, transmission, storage and search system and method

(57) A digital speech system processes acquired speech to determine speech element information, such as the phonemes in the speech, and associated the prosody information. This determined information is then encoded for transmission and/or storage. The encoded information is decoded to recover the speech element information and the prosody information which can be provided to a speech generator to construct a facsimile of the original speech. Also, the speech element and prosody information can be searched to locate words or phrases of interest in the speech.

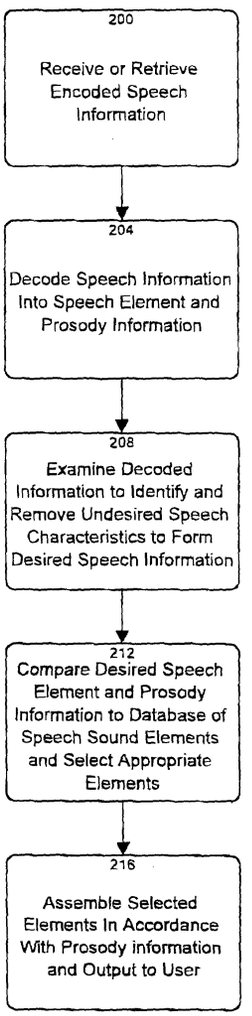


Fig. 6

EP 1 103 954 A1

## Description

### Field Of The Invention

**[0001]** The present invention relates to digital speech systems and methods. More particularly, the present invention relates to methods of acquiring, transmitting, storing and searching speech in a digital format.

### Background Of The Invention

**[0002]** Much research and effort has been directed to speech recognition by computer systems and/or to transmission of speech in digital formats. For example, IBM, Dragonsofl and others sell products for personal computers which allow a user to speak computer commands into a microphone and have the computer respond accordingly to the spoken commands. More advanced systems which have recently become available allow a user to dictate into the microphone and have the computer transcribe the spoken words into written text.

**[0003]** These systems operate by converting the analog signal produced by the microphone into a digital signal. The computer system then analyses the digital signals, typically employing a technique called Hidden Markov Modeling (HMM), in an attempt to extract content and meaning from digitized speech. When employed with telephony and restricted ranges of vocabulary, recent HMM systems have been able to achieve rates of accuracy approaching 90%, or better.

**[0004]** However, HMM-based systems suffer from several disadvantages. Specifically, HMM-based systems employ pattern matching and statistical models to assign words to the digitized speech. This can require significant amounts of processing power and such systems do not offer good accuracy when the vocabulary they must analyze is not very restricted. Further, these systems typically require configuration (training) to be performed by each user to learn the user's speech characteristics before a user can utilize the system. Even when trained, HMM-based systems can make errors. For example, in one test of an HMM-based system after a full training session, the phrase, "I am based in Toronto" was transcribed as "I embraced Toronto" and "Ian blazed Toronto". Further specific training was required before "based" could be recognized by the system.

**[0005]** It is also known to transmit or store voice data in digital form. For example, a compact disc can store a speech and telephone systems now commonly employ digital links over which the conversations are carried. More recently, voice conversations have been transmitted over digital packet networks and, it has recently become possible to conduct telephone conversations over the internet. In the intemet systems, the user speaks into a microphone at a personal computer and the output of the microphone is digitized, packetized and transmitted through the internet by the personal computer to a remote computer where it is received, converted back into an analog signal and output to the receiving user.

**[0006]** However, digital voice storage and transmission systems also suffer from disadvantages. One of the primary disadvantages is that, to obtain reasonable quality voice signals, the digitization process must have a relatively high sampling rate (e.g. a minimum of 6,500 samples per second and preferably 8,000 or 10,000) and thus digitized voice requires a relatively high amount of bandwidth (e.g. about 1,500 bits per second) for transmission. For example, Voxware, Inc., 305 College Road East, Princeton, New Jersey, USA sells a variety of digital voice codecs. These codecs employ 8kHz sampling of a 4kHz speech frequency and require between 1260 and 6400 bps (bits per second) for transmission, depending upon the desired quality and robustness of the transmission.

**[0007]** An additional disadvantage is that, if such digital voice transmissions are to be stored for later access, which can be required for call logging and other purposes, a large amount of storage space can be required. These bandwidth and/or storage requirements can prevent voice transmission from being provided as a feature for intemet multi-player games, etc.

**[0008]** Yet another disadvantage with prior art digital speech systems is that they do not provide an acceptable method of searching the speech for words or phrases of interest. At best, the prior art systems allow "overspeed" playback of the speech to allow a user to listen to the speech at a fast playback rate to attempt to locate portions of interest in the speech which can then be played back at the correct playback rate. At worst, these systems require that the user listen to the speech at correct playback rates to locate portions of interest. These search techniques leave much to be desired, especially if a long portion of speech is to be searched.

**[0009]** It is desired to have a system and method which allows relatively efficient acquisition, transmission, searching and storage of speech information in a digital form.

### Summary Of The Invention

**[0010]** It is an object of the present invention to provide a novel system and method of acquiring, transmitting, searching and/or storing speech information which obviates or mitigates at least one of the above-mentioned disadvantages of the prior art.

**[0011]** According to a first aspect of the present invention, there is provided a digital speech system, comprising:

speech element determination means operable on speech in a digital format to determine speech element information from said speech;

speech prosody determination means operable on said speech to determine prosody information from said speech;

encoding means to encode speech information

comprising said determined speech element information, said determined speech prosody information and timing information relating thereto in a digital form;
decoding means to decode said encoded speech information to obtain said determined speech element information, said determined speech prosody information and said timing information;
comparison means to compare said determined speech element information and said determined speech prosody information to a database of speech elements to select speech sound elements which correspond thereto; and
speech generating means to assemble said selected speech sound elements to construct a facsimile of said speech.

**[0012]** Preferably, the system further comprises speech acquisition means to acquire an analog electronic representation of the speech and digitization means to convert the analog representation of the speech to the digital format. Also preferably, the system further comprises an output means to output the facsimile to a user in an audible manner.

**[0013]** According to another aspect of the present invention, there is provided a method of acquiring and constructing digital speech, comprising the steps of:

examining digitized speech to determine speech element information relating to said speech;
examining said digitized speech to determine prosody information relating to said speech;
encoding speech information corresponding to said determined speech element information, said determined prosody information and timing information relating thereto;
receiving and decoding said encoded speech information to obtain said timing information, said determined speech element information and said determined prosody information;
comparing said decoded speech element information and prosody information to a database to select corresponding speech sound elements; and
assembling said selected speech sound elements to construct a facsimile of said speech.

## Brief Description Of The Drawings

**[0014]** Preferred embodiments of the present invention will now be described, by way of example only, with reference to the attached Figures, wherein:

Figure 1 shows a block diagram representation of a digital speech system in accordance with an embodiment of the present invention;
Figure 2 shows a flowchart of a digital speech acquisition process of Figure 1;
Figure 3 shows a flowchart of the speech generation process of Figure 1;
Figure 4 shows a schematic representation of a telephone-like communication system operating between two users connected to the internet in accordance with the present invention;
Figure 5 shows a schematic representation of a call center employing an embodiment of the present invention; and
Figure 6 shows a flowchart of a speech generation process in accordance with another embodiment of the present invention.

## Detailed Description Of The Invention

**[0015]** A digital speech system, in accordance with an embodiment of the present invention, is indicated generally at 10 in Figure 1. System 10 includes a transducer 12, such as a microphone or other suitable means to acquire speech in electronic form, a speech acquisition portion 14, as described below, a speech generation portion 16, also described below, and a speech output 18, such as a loudspeaker or other suitable means for presenting the constructed speech facsimile to a user or subsequent system. Connection 20 between speech acquisition portion 14 and speech generation portion 16 can be a connection, such as a data network, or can be a storage and retrieval system, such as a digital mass storage device which allows acquired speech to be stored and to subsequently be provided to speech generation portion 16.

**[0016]** As shown is Figure 2, speech acquisition portion 14 includes five principal functional blocks. In block 24, portion 14 acquires, via transducer 12, speech to be processed and in block 28 the acquired speech is digitized at an appropriate sampling rate. In block 32, the digitized speech is analyzed to determine speech element information from the acquired speech and in block 36 prosody information is determined from the acquired speech. As will be apparent to those of skill in the art, the functions of blocks 32 and 36 can be performed in a serial manner, as shown in the Figure, or can be performed in parallel if desired. In either case, the time information relating the determined speech elements to the determined prosody is part of the determined information. Once the speech element and prosody information is determined in blocks 32 and 36, this information is encoded by block 40 and the resulting encoded speech information can be transmitted and/or stored by connection 20, as indicated at block 44.

**[0017]** The acquisition and digitization of the speech at blocks 24 and 28 can be accomplished in any suitable manner as will occur to those of skill in the art and many conventional techniques to accomplish this are known. For example, in the environment of personal computers, it is not uncommon that speech can be acquired and digitized with an inexpensive microphone connected to an input port on a sound card, such as a SoundBlaster or compatible card. It is also contemplated that the func-

tions of blocks 24 and 28 can be performed in separately and/or at a different time. For example, the acquired and digitized speech can be provided to speech system 10 from a previously recorded CD ROM, etc.

**[0018]** The speech elements referred to with respect to block 32 can be any representative elements, such as allophones, diphones, phonemes, etc., as will occur to those of skill in the art. In a presently preferred embodiment of the invention, phonemes are determined from the acquired speech. Phonemes are a formal representation of phones (the distinct sounds in speech) and include vowels, semivowels, dipthongs and constants. Different languages and/or accents have different phonemes and, depending upon the phoneme definition standard employed and its implementation, there are between about forty and about sixty phonemes in English.

**[0019]** The analysis of speech, via digital processing, to determine phonemes is known and is, for example, discussed in, "Fundamentals of Speech Recognition: A Short Course", by Dr. Joseph Picone, Institute for Signal and Information Processing, Mississippi State University. This reference, the contents of which are incorporated herein by reference, is available at http://www.isip.msstate.edu/resources/index.html#journals.

**[0020]** Prosody, which is the information determined in block 36, represents pitch, timing and other information associated with speech. An example of prosody is the rising pitch at the end of a sentence to represent a question. Other examples of prosody include pauses to represent the ends of sentences, etc. Again, the determination of prosody information from speech is known. The paper, "Automatic Prosodic Analysis for Computer Aided Pronunciation Teaching", Paul Christopher Bagshaw, available at http://www.cstr.ed.ac.uk/pubs_by_year.html#theses provides a good description of the determination of prosody and the contents of this reference are also incorporated herein by reference.

**[0021]** Conventional speech element determination systems, for example phoneme determination systems, return a matrix for all possible phonemes indicating the probability that each particular phoneme was present. Word recognizer systems search for the matrix for the letter combinations that have the highest probability of being a word. Unlike these conventional speech element determination systems, the present invention can further employ these probabilities as a reference to tables of allophones, looking for sound similarities, overlap in sounds, and such, to assist in finding the most probable sound match. rather than the most probable word match. Matching on sounds, rather than words, results in a much better approximation of the way speech works, especially the way people mix and blend sounds in real speech.

**[0022]** Once the speech information, comprising the speech element and prosody information, has been determined in blocks 32 and 36, it is encoded at block 40. The information is encoded so that it can be effectively represented in a digital form. For example, if system 20 has sixty phonemes defined for it, for the English language, a set of sixty unique symbols can be defined to represent these phonemes and this symbol set can be represented by six bits of infonnation. As will be apparent, six bits are much less than the many bytes of information which would be required for a digitized sample of the spoken phoneme. One encoding scheme that can be employed is to use the Unicode values for the International Phonetic Alphabet to represent the phonetic information and this is the presently preferred encoding scheme for the determined speech elements.

**[0023]** Prosody information is also encoded in a digital form, for example as numbers representing the duration of speech elements and pitch changes. Again, the encoding of the determined prosody information can be accomplished with a relatively small symbol set and/or in a size efficient manner, compared to digitized samples of speech. As mentioned above, in the encoding, the time relationship between the speech elements and prosody information is maintained.

**[0024]** The information encoded at block 40 is stored or transmitted, through connection 20, at block 44.

**[0025]** The present invention allows a relatively low bandwidth method of transmitting speech information and/or a relatively efficient storage method for speech. As discussed above, prior art systems of which the inventor is aware either digitize speech for transmission and/or storage and reconvert that speech to an analog form for use, or perform a speech to text conversion. In contrast, the present invention determines information about the speech, the information being stored and/or transmitted in a manner that maintains the information without maintaining the actual speech. Specifically, the encoded determined information can be employed to construct a facsimile representation of the original speech or to otherwise interact and access the information of the original speech.

**[0026]** Construction of a facsimile representation of the original speech from the stored or transmitted encoded information can be accomplished in speech generation portion 16, as shown in Figure 3. As illustrated, speech generation portion 16 includes four principal functional blocks. In functional block 52 the encoded information is received at or retrieved by the speech generation process. In block 56, the encoded information is decoded to obtain the speech element information and prosody information. Next, in block 60, the speech information and related prosody information is compared to a database of speech sound elements available to the speech generator. Statistical modeling, or any other suitable selection process, is then employed to select appropriate available speech sound elements. In block 64, the selected speech sound elements are assembled in accordance with the prosody information and output to the user as a constructed facsimile of the original speech. As will be apparent to those of skill in the art, the speech sound elements can comprise digitized

speech samples and/or synthetic speech elements, such as those produced by a vocoder or by FM synthesis. In the latter case, the selected speech sound elements comprise appropriate instructions for the vocoder or FM synthesis system to create the desired sounds.

**[0027]** As will be apparent to those of skill in the art, the determination of speech elements and prosody information and the selection of speech sound elements for construction of facsimiles of the original speech can be less than completely accurate. However, as is well known, most individuals do not always speak in a perfectly correct manner and yet their speech is still understood by others. This understanding of 'flawed' speech is possible due to the redundancy of information in speech, including the context of the speech and the relationship between the speech elements and the prosody. The present invention takes advantage of the human ability to understand even flawed speech without undue effort and it has been determined by the present inventor that relatively many errors in the overall process of constructing a facsimile of the original speech are easily tolerated by most humans.

**[0028]** Phoneme determination, and speech determination in general, are statistical in nature and prone to errors and 'almost right' recognition. The present invention preferably takes advantage of the relationship in speech between phonemes and allophones to provide acceptable playback of speech. As mentioned above, English has between about forty and about sixty phonemes but there are many ways of pronouncing these phonemes, since they tend to be modified by dialect, preceding and following sounds, emotion, non-word speech components such as the pitch rise for questions, etc. In constructing a facsimile of the original speech, the present invention can use the speech relationship between phonemes and allophones to select the most probable phoneme matches by also considering the list of allophones. As the top matches tend to be similar, for example "f" and "v" might be the top recognition for the first phone in "fire", an allophone which is a close match to both can be found by examining the allophone list against the phoneme probabilities. Selecting the best phoneme match based on both the phoneme probability and the allophone probability allows the present invention to obtain matches very close to the original sound.

**[0029]** One use of the present invention. which is illustrated schematically in Figure 4, allows transmission of the encoded speech information to a remote location via a relatively low bandwidth data network. For example, a conversation can be conducted by two or more individuals over a packet network 100, such as the internet, with each party employing a computer system 104a, 104b such as a general purpose IBM compatible personal computer or a dedicated special purpose computer-based system. Computer system 104 operates to acquire the user's speech via a microphone 108a, 108b and to determine and encode the speech elements and prosody information as acquisition portion 14 of system 10. The encoded information is then transmitted through the intemet 100 to the other user. Further, each user's computer system 104 receives the encoded speech information from the other user or users and constructs a facsimile of the original speech as the speech generation portion 16 of system 10, as described above, which is then output to the user via a loudspeaker 112, headset or other suitable output device.

**[0030]** The relatively low bandwidth required to transmit the encoded speech information, which can also be compressed prior to transmission and de-compressed after reception if desired, allows the present invention to be employed to permit players of online games to have spoken communication with other players over the limited bandwidth of their connections to the internet over which the other game-related information must also be transmitted.

**[0031]** In addition to use for bandwidth efficient communications, the present invention can provide a number of other features, advantages and uses. For example, in the above-mentioned online gaming uses, the speech generation portion 16 at each user's computer system can employ digitized samples of'celebrity' voices as the speech sound elements. It is contemplated that a user can select a Darth Vader-like voice to be assigned to a first user and a HAL 9000-like voice to be assigned to a second user. Further, it is also contemplated that the speech generation portion 16 can include a predefined selection of celebrity of voices, and each first user can specify which of the available voices is to be used to reconstruct the first user's speech in the second user's speech generation portion 16.

**[0032]** The present invention can also be employed to make more efficient use of systems wherein bandwidth is expensive. For example, many corporations now operate call centers which, for labor cost and other reasons, are often located in locations which are distant from the customers who access the call center. While the final financial analysis still supports such distant call centers, much of the otherwise potential savings can be absorbed by the necessity that the corporation provide and pay for toll free access to the call center for the customers. While a call center would still have to fund toll free access, with the present invention it is possible to significantly reduce the total required bandwidth to reduce the corporation's costs. In particular, as shown in Figure 5, the corporation can provide sub-centers 160 which can be accessed locally by customers 164. Sub-centers 160, which can be personal computer systems, minicomputers, special purpose devices or any other suitable system as will occur to those of skill in the art, comprise the above-described speech acquisition portion 14 and speech generation portion 16 of system 10. Sub-centers 160 acquire, determine, encode and transmit speech information from customers 164 calling the sub-center 160 to call center 168 over a suitable communications link 166. Communications link 166 can be a leased data connection, a packet network, satellite link

or any other suitable data communications link, as will occur to those of skill in the art.

**[0033]** The transmitted speech information is received at a suitable call center computer system 168 at the call center. Call center computer system 168 can be personal computer systems, minicomputers, special purpose devices or any other suitable system as will occur to those of skill in the art, and also comprises the above-described speech acquisition portion 14 and speech generation portion 16 of system 10. When call center computer system 168 receives the transmitted speech information, speech generation portion 16 constructs a facsimile of the customer's original speech. The attendant 172 at the call center then responds to the inquiry and the attendant's speech is processed by the speech acquisition portion 14 of call center computer system 168 to determine, encode and transmit, over link 166, speech information to the appropriate sub-center 160 where a construction of a facsimile of the attendant's speech is created and provided to the customer 164 by sub-center 160.

**[0034]** In this manner, less bandwidth is required for communications link 166 between the sub-center 160 and call center computer system 168, thus reducing the corporation's costs. As is also indicated in the Figure, if desired and if sufficient processing capacity exists at sub-center 160, multiple calls can be simultaneously processed by sub-center 160 and speech information between sub-center 160 and call center computer system 168 can be appropriately multiplexed over the communications link 166.

**[0035]** In addition to reducing bandwidth requirements or making more efficient use of available bandwidth, the present invention also provides a process whereby space efficient storage of speech can be accomplished. In this specific case of a call center, every conversation between customers 164 and attendants 174 can be stored, in encoded speech information form, and requires less storage capacity than other conventional methods. Further, as discussed below, speech stored with the present invention can be effectively searched for the occurrence of words or phrases of interest.

**[0036]** As an additional advantage of the present invention, speech generation portions 16 employed at sub-centers 160 can employ a preferred voice. For example, the corporation for which the call center is operated can have a celebrity corporate spokesperson and the celebrity's voice can be employed in the sub-center speech generation portions 16.

**[0037]** An additional feature of the present invention is that it can be used to alter or remove cultural or regional accents and speech features of the speakers. One of the disadvantages with operating call centers distal customers is that accents and other speech features of the call center attendants can emphasize or otherwise interfere with the relationship between the attendant and the customer. For example, a customer from the US mid-west may be uncomfortable in speaking with an attendant in a call center in Ireland, who has a noticeable Irish accent. In the past, this has necessitated that call centers attempt to hire or train attendants without strong accents or other undesired identifying speech characteristics.

**[0038]** Figure 6 shows the principal functional blocks of a speech generation process in accordance with another embodiment of the present invention. In this embodiment, as the speech generation process constructs a facsimile of the original speech from determined speech elements and prosody, it is possible to identify and reduce, or even eliminate, accents and other undesired identifying speech characteristics. The speech generation process of Figure 6 is similar to that of Figure 3, discussed above. In block 200, encoded speech information is received via a communications link or is retrieved from a storage device. In block 204, the encoded information is decoded to obtain the speech elements and prosody information. In block 208, the decoded information is examined, using statistical models, pattern matching or other appropriate techniques, to recognize undesired speech characteristics and to substitute information into the speech element information and prosody information to reduce or remove those identified undesired characteristics to form modified speech information. In block 212, the modified speech information is used to select speech sound elements and in block 216 a constructed modified facsimile of the original speech is output to the user.

**[0039]** As will be apparent to those of skill in the art, the recognition and substitution process to remove undesired speech aspects can be performed in the speech acquisition portion 14, prior to transmission or storage of the speech information, or in the speech generation portion 16, as discussed above.

**[0040]** Another feature of the present invention is the ability to facilitate the location of desired phrases or words in speech. In many applications it is desired to have the ability to search stored speech for the occurrence of particular words or phrases. For example, an attorney may wish to locate a phrase in an audio recording of the testimony of a witness, which can be several hours in length. Presently, such a search can be conducted only by either manually or automatically transcribing the testimony to text and then performing conventional text searching operations or by listening to the testimony recording. Neither of these approaches is satisfactory in many circumstances, with manual transcription being expensive, automatic transcription being subject to errors and listening to the recording being time consuming.

**[0041]** With the present invention, the speech to be searched can be processed, with a speech acquisition process similar to that of Figure 3, to determine encoded speech information. Speech can either be stored only in encoded speech element form, such as for the call center call logging discussed above, or can be stored in

both an encoded speech element form and an analog or digitized form. In either case, the determined speech elements are searched to locate words or phrases of interest.

**[0042]** When it is desired to locate the occurrence of a word or phrase of interest, the word or phrase is processed by the speech acquisition process to determine encoded speech information for the word or phrase. The search is then performed on the speech by comparing the stored determined speech element and prosody information with the speech element and prosody information for the search word or phrase. A suitable pattern matching or statistical algorithm can be employed to locate high probability matches between the speech information and the search information. For applications where the speech has only been stored as encoded speech information, such as the above-mentioned call logging for a call center. the identified matches are provided to a speech generation process, as described above, to let the searcher hear a construction of the appropriate portions of the speech. For applications where the original speech, in analog or digitized form, has also been stored, the time information related to the located matches can be provided so that the user can directly access the appropriate times in the original speech. As will be apparent to those of skill in the art, the above-mentioned processes for removing accents and undesired speech characteristics can also be employed with this search process to augment the ability to locate phrases or words when the searcher and the searchee have different speech characteristics. As will also be apparent, the stored determined speech information can be indexed, in any suitable manner as will occur to those of skill in the art, to enhance the speed at which searches of large datasets can be performed.

**[0043]** The present invention can also be employed in voice transcription systems. For example, a user can provide dictation which is processed to obtain encoded speech information, as described above, which can be stored in an efficient manner. The encoded speech can be used to construct a facsimile of the original speech to the user, thus providing a spoken memo pad-like application, or can be subsequently transcribed to text. In the former case, digitized samples of the voice of the user can be user to reconstruct the speech, if desired. In the latter case, the transcription can be performed by humans, in a conventional manner, or can be performed by speech to text transcriptions systems, such as the above-mentioned Dragonsoft and other systems.

**[0044]** One of the perceived advantages of the present invention, when used with speech to text transcription systems, is that as the speech information is stored, it can be examined in a non-real time manner. This allows multiple passes to be performed by the speech to text transcription system. Also, as mentioned above, accents or other speech characteristics which can confuse a speech to text transcription system, can be removed or mitigated by the present invention and this can reduce or eliminate lengthy training of the transcription system by the user.

**[0045]** As will be apparent to those of skill in the art, the present invention can also be employed with voice messaging type applications. For example, encoded speech information can be emailed to a recipient, who will then provide the speech information to a speech generation process, which can be a plug-in component for the email program, to hear a constructed facsimile of the original speech.

**[0046]** In another aspect, the present invention can provide voice recognition of users. In this embodiment, several base samples of the user's speech are processed to obtain speech element and prosody base information for that user and this information is stored in a database. To subsequently identify that user, the user is requested to speak to provide a test sample of their speech to the system. The test sample is processed to obtain speech element and prosody information from the test sample and the obtained information is compared to the base information previously determined for the user. A statistical comparison is then performed between the information for the base samples and the test sample to determine the degree of correspondence therebetween. Unlike prior art voice printing systems, the present invention is less sensitive to differences in the microphone and environment with which the base and the test samples are acquired and does not require that the user repeat the particular base samples originally provided to the system.

**[0047]** The above-described embodiments of the invention are intended to be examples of the present invention and alterations and modifications may be effected thereto, by those of skill in the art, without departing from the scope of the invention which is defined solely by the claims appended hereto.

## Claims

**1.** A digital speech system, comprising:

speech element determination means operable on speech in a digital format to determine speech element information from said speech;

speech prosody determination means operable on said speech to determine prosody information from said speech;

encoding means to encode speech information comprising said determined speech element information, said determined speech prosody information and timing information relating thereto in a digital form;

decoding means to decode said encoded speech information to obtain said determined speech element information, said determined speech prosody information and said timing information;

comparison means to compare said determined speech element information and said determined speech prosody information to a database of speech elements to select speech sound elements which correspond thereto; and speech generating means to assemble said selected speech sound elements to construct a facsimile of said speech.

**2.** The digital speech system of claim 1 further comprising:

speech acquisition means to acquire an analog electronic representation of said speech; and digitization means to convert said analog representation of said speech to said digital format.

**3.** The digital speech system of claim 1 further comprising:

an output means to produce an output of said facsimile in a manner audible to a user.

**4.** The digital speech system of claim 1 wherein said comparison means further comprises recognition means to identify undesired speech characteristics, said selection of speech sound elements being performed to reduce the presence of said identified undesired speech characteristics in said facsimile.

**5.** The digital speech system of claim 1 further comprising search means operable to receive an input representing a word or phrase of interest and to examine said determined speech element information to locate occurrences of said word or phrase therein.

**6.** A method of acquiring and constructing digital speech, comprising the steps of:

(i) examining digitized speech to determine speech element information relating to said speech;
(ii) examining said digitized speech to determine prosody information relating to said speech;
(iii) encoding speech information corresponding to said determined speech element information, said determined prosody information and timing information relating thereto;
(iv) receiving and decoding said encoded speech information to obtain said timing information, said determined speech element information and said determined prosody information;
(v) comparing said decoded speech element information and prosody information to a database to select corresponding speech sound elements; and
(vi) assembling said selected speech sound elements to construct a facsimile of said speech.

**7.** The method of claim 6 further comprising the steps of:

acquiring an electronic representation of speech in an analog form; and
digitizing said analog representation of speech to obtain said digitized speech for step (i),

**8.** The method of claim 6 wherein step (v) further comprises comparing said decoded speech element information to a predefined database of undesired speech characteristics and selecting speech sound elements which reduce said undesired speech characteristics is said facsimile.

**9.** The method of claim 6 further comprising the step of receiving from a user a word or phrase of interest and search said determined speech element information and said determined prosody information to locate occurrences of said received word or phrase and to identify said locations to said user.

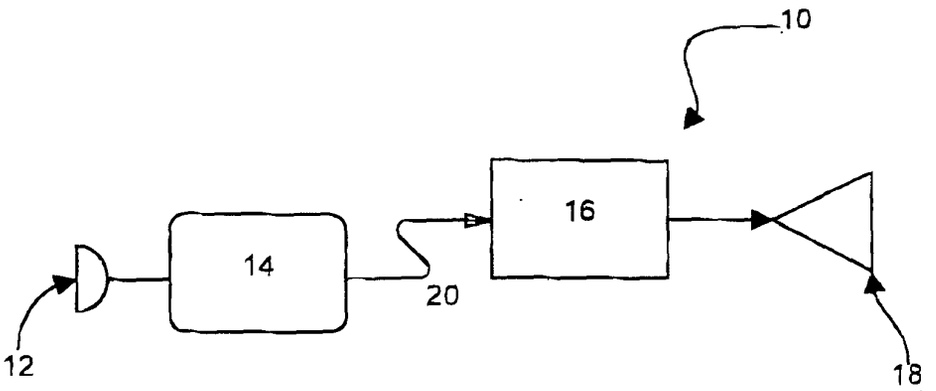
10
16
14
20
12
18

Fig. 1

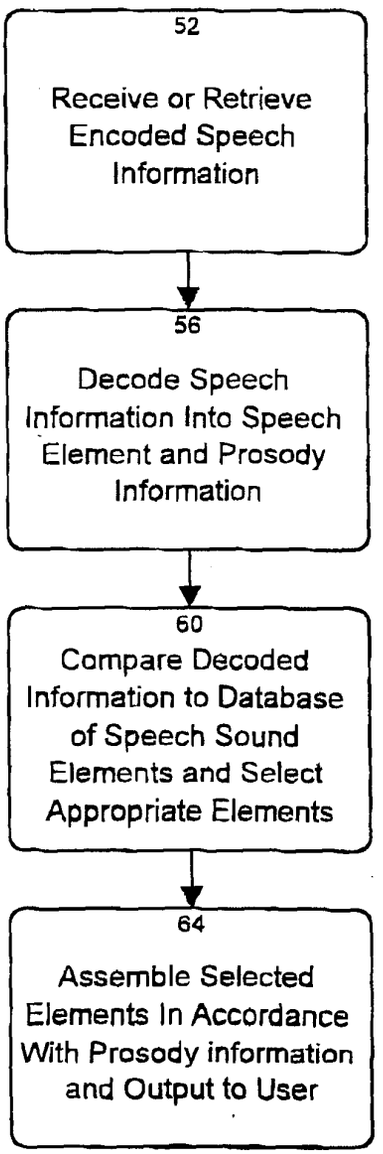
52
Receive or Retrieve Encoded Speech Information
56
Decode Speech Information Into Speech Element and Prosody Information
60
Compare Decoded Information to Database of Speech Sound Elements and Select Appropriate Elements
64
Assemble Selected Elements In Accordance With Prosody information and Output to User

Fig. 3

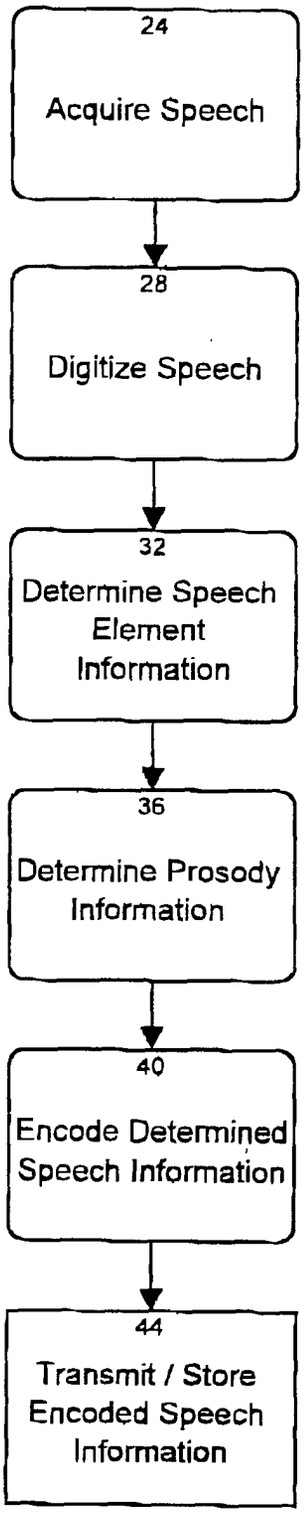
24
Acquire Speech
28
Digitize Speech
32
Determine Speech Element Information
36
Determine Prosody Information
40
Encode Determined Speech Information
44
Transmit / Store Encoded Speech Information

Fig. 2

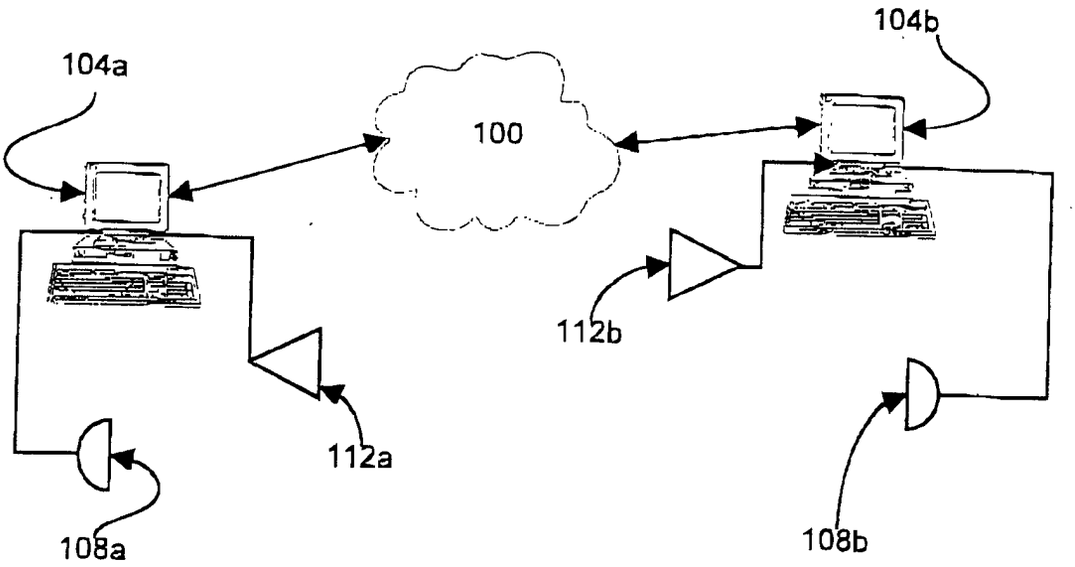


Fig. 4

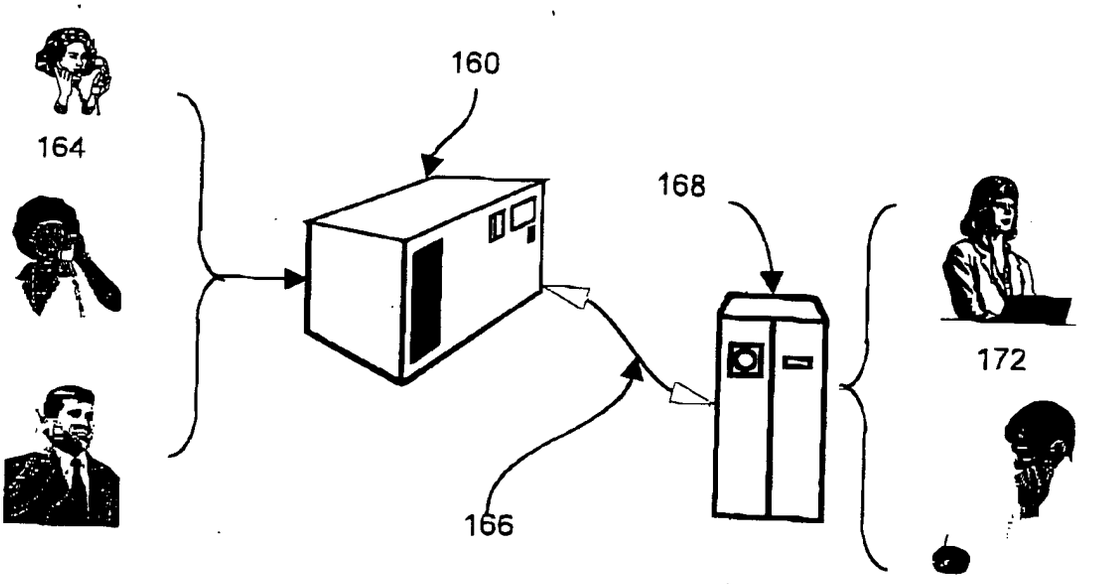


Fig. 5

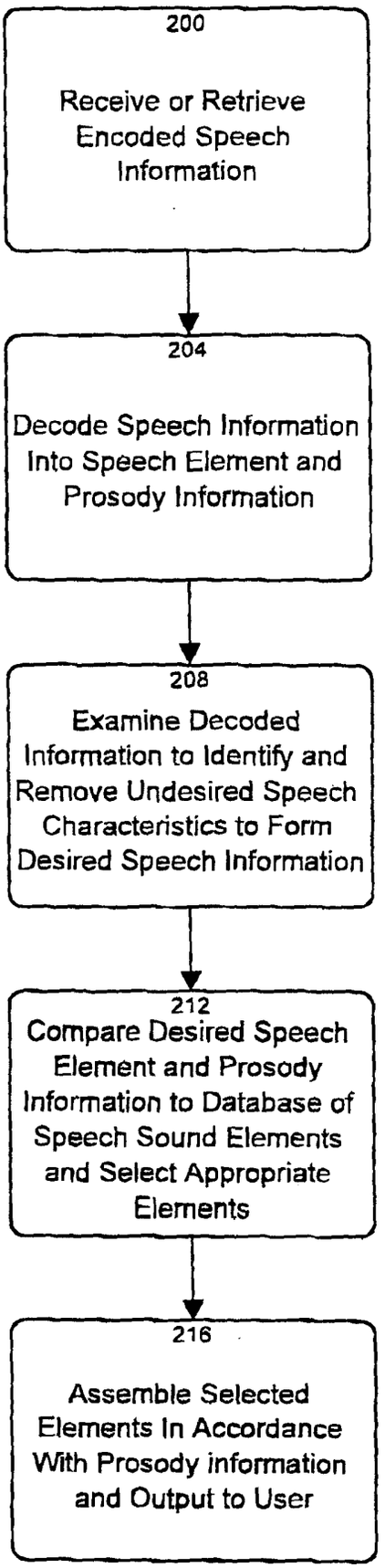
200
Receive or Retrieve Encoded Speech Information
204
Decode Speech Information Into Speech Element and Prosody Information
208
Examine Decoded Information to Identify and Remove Undesired Speech Characteristics to Form Desired Speech Information
212
Compare Desired Speech Element and Prosody Information to Database of Speech Sound Elements and Select Appropriate Elements
216
Assemble Selected Elements In Accordance With Prosody information and Output to User

Fig. 6

European Patent Office

# EUROPEAN SEARCH REPORT

Application Number
EP 99 20 3979

## DOCUMENTS CONSIDERED TO BE RELEVANT

| Category | Citation of document with indication, where appropriate, of relevant passages | Relevant to claim | CLASSIFICATION OF THE APPLICATION (Int.Cl.7) |
|---|---|---|---|
| X<br>Y | US 5 933 805 A (BOSS ET AL)<br>3 August 1999 (1999-08-03)<br>* abstract; figures 4,5 *<br>* column 2, line 59 - column 3, line 37 *<br>* column 4, line 32-67 *<br>* column 5, line 52-59 *<br>* column 7, line 32-38 * | 1-4,6-8<br>5,9 | G10L19/00 |
| Y | US 5 649 060 A (ELLOZY ET AL)<br>15 July 1997 (1997-07-15)<br>* abstract; figure 3 * | 5,9 | |
| A | GB 2 332 841 A (MOTOROLA LTD)<br>30 June 1999 (1999-06-30)<br>* abstract; figure 1 *<br>* page 1, line 24 - page 2, line 4 *<br>* page 3, line 28 - page 4, line 13 * | 1,4,6,8 | |
| | | | TECHNICAL FIELDS SEARCHED (Int.Cl.7) |
| | | | G10L |

The present search report has been drawn up for all claims

| Place of search | Date of completion of the search | Examiner |
|---|---|---|
| THE HAGUE | 6 April 2000 | Quélavoine, R |

CATEGORY OF CITED DOCUMENTS

X : particularly relevant if taken alone
Y : particularly relevant if combined with another document of the same category
A : technological background
O : non-written disclosure
P : intermediate document

T : theory or principle underlying the invention
E : earlier patent document, but published on, or after the filing date
D : document cited in the application
L : document cited for other reasons

& : member of the same patent family, corresponding document

EPO FORM 1503 03.82 (P04C01)

## ANNEX TO THE EUROPEAN SEARCH REPORT ON EUROPEAN PATENT APPLICATION NO. EP 99 20 3979

This annex lists the patent family members relating to the patent documents cited in the above-mentioned European search report. The members are as contained in the European Patent Office EDP file on
The European Patent Office is in no way liable for these particulars which are merely given for the purpose of information.

06-04-2000

| Patent document cited in search report | Publication date | Patent family member(s) | Publication date |
|---|---|---|---|
| US 5933805 A | 03-08-1999 | NONE | |
| US 5649060 A | 15-07-1997 | DE 69422466 D | 10-02-2000 |
| | | EP 0649144 A | 19-04-1995 |
| | | JP 2986345 B | 06-12-1999 |
| | | JP 7199379 A | 04-08-1995 |
| GB 2332841 A | 30-06-1999 | NONE | |

EPO FORM P0459

For more details about this annex : see Official Journal of the European Patent Office, No. 12/82