

(19) World Intellectual Property Organization
International Bureau



(43) International Publication Date
11 May 2006 (11.05.2006)

PCT

(10) International Publication Number
WO 2006/050238 A1

(51) International Patent Classification:
G10L 13/06 (2006.01)

(21) International Application Number:

PCT/US2005/039207

(22) International Filing Date: 28 October 2005 (28.10.2005)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:

60/622,838 28 October 2004 (28.10.2004) US

(71) Applicant (for all designated States except US): **VOICE SIGNAL TECHNOLOGIES, INC.** [US/US]; 150 Presidential Way, Woburn, MA 01801-1100 (US).

(72) Inventors; and

(75) Inventors/Applicants (for US only): **EDGINGTON, Michael** [US/US]; 15 Sarah Drive, Bridgewater, MA 02324 (US). **GILLICK, Laurence, S.** [US/US]; 55 Esty Farm Road, Newton, MA 02459 (US). **ZLOKARNIK, Igor** [DE/US]; 24 Emerson Street, Natick, MA 01760 (US).

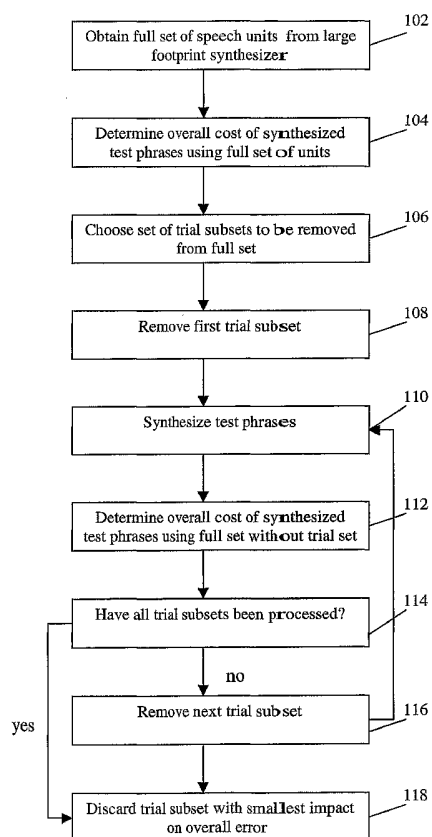
(74) Agents: **PRAHL, Eric** et al.; WILMER CUTLER PICKERING HALE AND DORR LLP, 60 State Street, Boston, Massachusetts 02109 (US).

(81) Designated States (unless otherwise indicated, for every kind of national protection available): AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BW, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KM, KN, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, LY, MA, MD, MG, MK, MN, MW, MX, MZ, NA, NG, NI, NO, NZ, OM, PG, PH, PL, PT, RO, RU, SC, SD, SE, SG, SK, SL, SM, SY, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, YU, ZA, ZM, ZW.

(84) Designated States (unless otherwise indicated, for every kind of regional protection available): ARIPO (BW, GH, GM, KE, LS, MW, MZ, NA, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HU, IE, IS, IT, LT, LU, LV, MC, NL, PL, PT,

[Continued on next page]

(54) Title: CODEC-DEPENDENT UNIT SELECTION FOR MOBILE DEVICES



(57) Abstract: A method of extracting a subset of speech units from a larger set of speech units for use by a speech synthesizer in synthesizing speech, wherein the speech units are stored in a compressed encoded representation that was generated by a codec, the method comprising: selecting members of the subset of speech units based on an overall cost associated with using the speech synthesizer to synthesize a test set of speech, wherein the overall cost includes at least one error introduced by using the codec to decode the stored representations of the speech units; and storing the selected subset of speech units on a speech-enabled device.

WO 2006/050238 A1



RO, SE, SI, SK, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

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

Published:

- *with international search report*
- *before the expiration of the time limit for amending the claims and to be republished in the event of receipt of amendments*

Codec-Dependent Unit Selection for Mobile Devices

[0001] This application claims the benefit of U.S. Provisional Application No. 60/622,838, filed October 28, 2004, and incorporated herein by reference.

Technical field of the invention

[0002] This invention relates to speech synthesis, embedded software, and mobile devices.

Background

[0003] Speech synthesizers are used to transform text into speech. Modern speech synthesis systems are designed using “unit selection”, where short snippets of speech are selected from a database of utterances of a speaker or speakers. The snippets may then be reassembled using an algorithm to approximately recreate the sound of any particular word or utterance in the language. Classical unit selection depends on two cost functions, one, called the target cost, for the differences between features associated with a sound in a word and the relevant features of the synthesized sound in that word, and a second cost called the continuity cost, associated with the weighted sum of the differences of features at the junctures of the selected units in synthesized utterances. Units are selected so that the synthesizer database contains units that minimize these two costs, measured over a training set, and given a dictionary size, synthesis rules, and weights for the two cost functions.

Summary

[0004] The described embodiment has one or more of the following advantages. The described methods take into account the effect of distortion of synthesized units introduced by compression of the units by a vocoder. The described approach chooses a small subset of units for a small footprint synthesizer, basing the selection not only on the target cost and continuity cost from the training set, but also on an analysis of the fidelity of the acoustic signal reproduced by the codec.

[0005] In general, in one aspect, the invention features a method of extracting a subset of speech units from a larger set of speech units for use by a speech synthesizer in synthesizing speech when the speech units are stored in a compressed encoded representation that was generated by a codec. The method involves: selecting

members of the subset of speech units based on an overall cost associated with using the speech synthesizer to synthesize a test set of speech, wherein the overall cost includes an error introduced by using the codec to decode the stored representations of the speech units; and storing the selected subset of speech units on a speech-enabled device.

[0006] Other embodiments include one or more of the following features. For each member of a plurality of members of the larger set of speech units, the overall cost associated with using the speech synthesizer to synthesize the test set of speech from a population of speech units that excludes that member is computed. Speech units are selected based on the overall costs computed for the plurality of members. The error introduced by the codec includes at least one of an error associated with decoding individual stored representations of the speech units and an error associated with a continuity between successive decoded stored representations of the speech units. The overall cost includes a weighted sum of a target cost, a continuity cost, and an error introduced by using the codec to decode the stored representations of the speech units. Selecting the weights involves optimizing a quality of the synthesized test set of speech by minimizing the overall cost as defined by the selected weights. Determining the quality of the synthesized set involves at least one of an empirical comparison of the synthesized speech with the original test set by a human listener and an objective comparison. Some systems pre-assign speech units to clusters based on linguistic features of the speech unit. In such systems, a method of selecting the subset of units involves selecting at least one unit from each cluster. Selecting the units from within a cluster is based at least in part on a measure of how well the unit represents the cluster acoustically. Determining the size of the extracted subset of speech units depends on a storage constraint on the device on which the speech synthesizer resides.

Brief Description of the Figures

[0007] Fig. 1 is a flow diagram showing an algorithm for selecting a subset of units from a set of units, taking into account the effect of the vocoder.

Fig. 2 shows a high-level block diagram of a mobile phone incorporating a speech synthesizer.

Detailed Description

[0008] As described above, traditional methods of designing a large vocabulary unit-selection speech synthesizer are based on the minimization of error in speech synthesis systems that store uncompressed speech snippets. The error is composed of two parts: target cost and continuity cost. Target cost corresponds to the mismatch between the snippet sounds and the sounds of the target word fragments that the snippet is attempting to match. The target cost C^t for using speech unit u_i to synthesize a target speech fragment t_i can be defined as the weighted sum of differences of relevant features:

$$C^t(t_i, u_i) = \sum_{j=1}^p w_j^t C_j^t(t_i, u_i).$$

[0009] The relevant features include phonetic, metrical, and prosodic context.

The continuity cost represents mismatches occurring at the joins between successive snippets:

$$C^c(u_{i-1}, u_i) = \sum_{k=1}^q w_k^c C_k^c(u_{i-1}, u_i).$$

[0010] Traditional methods of unit selection select the units that minimize the sum of the target cost and the continuity costs. This process is described in Black, A. (2002) *Perfect Synthesis for all of the people all of the time*, Keynote, IEEE TTS Workshop 2002, Santa Monica, CA, <http://www.cs.cmu.edu/~awb/papers/IEEE2002/allthetime/node1.html>, which is incorporated herein in its entirety.

[0011] The units correspond to any speech fragment that can be concatenated to synthesize speech. Typically, units are half-phones or a multiple of half-phones.

[0012] Some speech-enabled devices, such as mobile computers, digital cell phones, or PDAs, impose stringent storage limits on resident applications. Small footprint synthesizers are required for such mobile devices. To meet these constraints, speech synthesizers use speech snippets that are compressed using a lossy compression scheme that reduces the fidelity of snippet synthesis. In these systems, the particular coding scheme for the stored speech can affect the intelligibility of the

resultant synthesized speech, as the codings are non-linear and affect different speakers in different ways.

[0013] The described embodiment enhances the selection of compressed snippets for a small footprint synthesis system by including the effect of the compression on the synthesized speech in sub-selecting the units from a large vocabulary unit-selection speech synthesizer. Compression affects both target and continuity costs. The encoder cost described below refers to encoder impact on the target cost. Also described below is a further enhancement in which the encoder impact on continuity cost is considered.

[0014] Standard methods of optimizing the functionality within a limited footprint involve compression of the speech snippets using a resident codec, such as the Enhanced Variable Rate Codec (EVRC) as specified by the TIA/EIA/IS-127 Interim Standard of January 1997, which is incorporated herein in its entirety. In such systems, the synthesized speech is subject not only to errors corresponding to the target cost and the continuity cost, but also to errors introduced by the imperfections of the codec compression and decompression. These errors can interact in a non-linear fashion with the target and continuity costs to produce a significant impact on the intelligibility of synthesized speech.

[0015] Speech synthesizers typically work with a set of a few tens or hundreds of thousands of speech units. The storage constraints of mobile digital devices often require that this set be reduced by a factor of ten or more, to only a few thousand units. This radical pruning of units makes it critical that the remaining units optimally represent the speech to be synthesized in the actual operating environment in which codec-induced distortion affects the synthesized speech.

[0016] In order to optimize the selection of snippets with respect to the actual codec-processed speech heard by a user, a third error, called the encoder cost, is introduced. The encoder cost takes account of the codec-induced distortions. A method of determining the encoder cost is described below.

[0017] In the described embodiment the overall cost for synthesizing the training set using a particular unit subset is the weighted sum of all three component costs: the traditional target and continuity costs, and the new component – the encoder cost:

$$C = w_t * C^t + w_c * C^c + w_e * C^e$$

[0018] Minimizing the overall cost in order to select the optimal subset of units represents a compromise between the three component costs. The system trades off local phonetic accuracy (low target cost), smoothness (low continuity cost), and good representation by the vocoder (low encoder cost).

[0019] Each component cost is typically formulated separately and may itself be a weighted sum of costs. For example, the target cost can be expressed in terms of a weighted sum of costs introduced by relevant features of speech, such as phonetic, metrical, and prosodic context. The encoder cost may be approximated to the encoder's internal measurement of coding error, which it determines as follows. For each frame, the encoder takes measurements of the speech signal in several dimensions in acoustic space, models these with a set of parameters for the encoder's speech model, and then stores a quantized representation of these parameters. The parameters are chosen by using an analysis-by-synthesis procedure in which the speech synthesized using the parameters is compared to the natural speech and an error is computed. The model parameters are adjusted until this error is reduced to an acceptable level. Thus an explicit error measure of the difference between the codec's model prediction and the natural speech is available in the form of the codec internal error measure.

[0020] In an alternative approach, the system computes the model parameters directly and quantizes them in order to represent them efficiently in the bitstream. The system typically quantizes the set of parameters by comparing the parameters to predetermined sets of parameter values, each of which is represented by a unique codeword. The system picks the predetermined set of parameters closest to the computed model parameters and determines an error between the computed parameters and the selected predetermined (i.e., quantized) set of parameters. This error is a measure of the quantization error, and can also be used to estimate the encoder cost.

[0021] The system uses empirical or objective approaches to determine the weights w_t , w_c , and w_e . The empirical approach uses judgments by human subjects about the quality of the speech. For example, one or more skilled listeners, such as the system developers themselves, adjust weights manually to determine what setting

sounds best. In a more controlled subjective evaluation, a set of listeners are given speech stimuli and asked to make a scaled quality judgment. This approach might not be practical for searching a large space of potential weights, but can be used when choosing amongst a small set of candidate configurations for a system. The optimal set of weights is that which causes the system to synthesize the “best” sounding speech when it minimizes overall cost as defined by the optimally weighted sum of the three component costs.

[0022] The objective approach uses a perceptually motivated measurement of speech quality that compares the synthesized speech with the natural speech. In one example, a system calculates the Mel-scaled cepstral coefficient (MFCC) for every frame of the natural and synthesized utterances so that the boundaries align, and determines the root mean square difference between the MFCC vectors.

[0023] Inclusion of the encoder cost in the error estimation will, in general, cause a different set of snippets to be selected for each codec. Thus a CDMA phone device will have an optimized speech synthesizer consisting of a different subset of snippets from those selected for a GSM phone, since the two phones use different codecs. In each case, the intelligibility of the synthesizer will be subjected to the described optimization process, given the training set, footprint constraints, and the specified codec.

[0024] Put another way, the selection of snippets is responsive to the codec's impact on different kinds of speech. For example, if a particular codec represents a certain kind of speech poorly, the optimized system allocates more units to cover this kind of speech so as to compensate for the codec's shortcomings. Conversely, other types of speech that are well represented by the codec will be allocated fewer units.

[0025] The unit selection process can be further enhanced by including in the overall cost a continuity cost component that also takes into account the performance of the codec:

$$C = w_t * C^t + w_{ce} * C^{ce} + w_e * C^e .$$

[0026] All codecs of interest use the inherent correlation between consecutive frames of speech to reduce the amount of information that needs to be represented in the encoded bitstream. If the codec's measurement of correlation between frames is high at the first frame of a specific unit, u_n , then the overall system will be sensitive to

any change between the acoustic details of the original context (the preceding unit u_{n-1}) and any other candidate u_m . The measured correlation in the original context and the difference in acoustic features between the original context and the new candidate context provides an estimate of the contiguity cost between any arbitrary pair of units in the inventory. For example, consider a speech utterance in the training set: "the cat." If speech units correspond to half phones, then the codec would measure a high degree of correlation between the two successive units from the "e" sound in "the." By contrast, there is a big change between the two successive units representing the second half of the "e" sound and the first part of the "c" sound at the beginning of "cat", so the codec would measure a low correlation. The degree of correlation will be implicitly represented in the encoded bitstream for the second unit of each pair.

During synthesis, the codec maintains the assumption of high correlation for the first pair, so that any difference between the original unit for the first part of "e", and a new unit in the reduced inventory, has a large effect on the subsequent unit, i.e. the second part of the "e" sound. Where the codec expects low correlation, as in the latter pair, the differences have a more localized effect in the first of the pair. Therefore the quality of the synthesized speech is less sensitive to differences between the original context and synthetic context when the codec is anticipating a larger discontinuity.

[0027] The optimal subset of speech units is selected from the population of speech units as illustrated in Fig. 1. The system obtains the full set of speech units from a large footprint synthesizer (102) and determines the overall cost of synthesizing a set of test phrases (104). In general the test phrases will be different from the training set and natural speech will not be available for the test phrases. The optimization process uses test phrases that correspond as closely as possible to the body of speech that the speech synthesizer is expected to be called upon to synthesize. The system determines what trial subsets to eliminate (106) in the selection process, as described below. The trial subsets can be individual speech units, but the number of iterations in the process can be reduced if more than one unit is eliminated in each trial subset. According to one heuristic, such acceleration of the optimization process is especially effective when the elements of the trial subset include syllable or word-length sequences of units that appear contiguously in the training set.

[0028] The unit selection process proceeds by removing the first trial subset (108), synthesizing test phrases using the remaining units (110), and determining the overall cost of the synthesized test phrases (112) as described above. The process is

repeated, each time omitting a different trial subset from the set of speech units (114, 116), synthesizing the test phrases, and determining the overall cost. The trial subset which, when omitted, make the smallest contribution to the increase in overall cost is discarded (118). The above steps are repeated, each iteration reducing the size of the remaining set of speech units, until the remaining speech units fit within the imposed storage footprint limit. Typically this limit corresponds to an allocation of memory on a speech-enabled mobile device. This subset of units is the optimal subset since it makes the largest contributions to overall cost reduction.

[0029] An alternative approach makes use of clustering analysis. Many large vocabulary speech synthesizers employ clusters while synthesizing speech in real time in order to reduce the unit search space based on the speech's linguistic context. An illustrative clustering technique is described in M. W. Macon, A. E. Cronk, and J. Wouters, *Generalization and discrimination in tree-structured unit selection*, in Proceedings of the 3rd ESCA/COCOSDA International Speech Synthesis Workshop, November 1998, and in Hunt, A.J. and Black, A.W., *Unit Selection in a Concatenative Speech Synthesis System Using a Large Speech Database*, in Proc. ICASSP-96, May 7-10, Atlanta, GA, which are incorporated herein in their entirety. In cluster analysis, a speech unit is pre-classified and placed into a cluster with other units that have similar linguistic features. Linguistic features are features that are derived from the text, such as the phoneme sequence, lexical stress, syllable structure, proximity to the start/end of the current phrase, the likelihood of significant pitch movement in the syllable, and utterance type (such as a question or a statement). Clusters are linked by decision trees that determine cluster sequence probabilities. For example, a single cluster might be used for the first half of all primary stressed "ee" sounds when followed by a nasal, preceded by a "t" and not in the last word of a phrase. When using clusters, a synthesizer determines the subsequent cluster using the decision tree. The synthesizer searches for the best unit only within that subsequent cluster. Since only a small fraction of the available units lie within a given cluster, the search space is radically pruned.

[0030] Since cluster analysis places similar units within the same cluster, an effective unit selection approach is to use clusters as the "first cut" for selecting units for the small footprint synthesizer by selecting at least one representative unit from each cluster. To determine which units within a cluster best represent the cluster, the

system evaluates the cluster units acoustically using, for example, their MFCC, and discards the acoustic outliers. The remaining units are then evaluated against the applicable codec and their encoder cost is determined. Units within each cluster are then selected based on an overall utility parameter that is a combination of their encoder cost and of how well they represent the cluster acoustically.

[0031] A minimum of one unit from each cluster is selected, but more can be selected until the footprint constraints are reached. In one approach, the number of units selected from a given cluster depends on the frequency that the cluster is selected over a representative set of test phrases.

[0032] In the described embodiment, the speech-enabled device is a cellular phone 200, such as is illustrated by the high-level functional block diagram of Fig. 2. The phone 200 is a Microsoft PocketPC-powered phone which includes at its core a baseband DSP 202 (digital signal processor) for handling the cellular communication functions (including for example voiceband and channel coding functions) and an applications processor 204 (e.g., Intel StrongArm SA-1110) on which the PocketPC operating system runs. The phone supports various applications, such as GSM voice calls, SMS (Short Messaging Service) text messaging, wireless email, and desktop-like web browsing along with more traditional PDA features.

[0033] The transmit and receive functions are implemented by an RF synthesizer 206 and an RF radio transceiver 208 followed by a power amplifier module 210 that handles the final-stage RF transmit duties through an antenna 212. An interface ASIC (application-specific integrated circuit) 214 and an audio Codec 216 provide interfaces to a speaker, a microphone, and other input/output devices provided in the phone such as a numeric or alphanumeric keypad (not shown) for entering commands and information. DSP 202 uses a flash memory 218 for code store. A Li-Ion (lithium-ion) battery 220 powers the phone and a power management module 222 coupled to DSP 202 manages power consumption within the phone. Volatile and non-volatile memory for applications processor 204 is provided in the form of SDRAM 224 and flash memory 226, respectively. This arrangement of memory is used to hold the code for the operating system, the code for customizable features such as the phone directory, and the code for any applications software that might be included in the phone, including the compressed speech units and speech synthesizer described above.

[0034] The visual display device for the phone includes an LCD driver chip 228 that drives an LCD display 230. There is also a clock module 232 that provides the clock signals for the other devices within the phone and provides an indicator of real time.

[0035] All of the above-described components are packaged within an appropriately designed housing 234. Since the phone described above is representative of the general internal structure of a number of different commercially available phones and since the internal circuit design of those phones is generally known to persons of ordinary skill in this art, further details about the components shown in Fig. 2 and their operation are not being provided and are not necessary to understanding the invention.

[0036] In general, the speech enabled device would not have to be a cellular phone at all, but would possess the functionality of synthesizing speech from an internally stored coded set of speech units, decoding the speech units, and synthesizing speech from the units. For example a laptop computer or PDA having a speaker, and appropriate software to generate speech from speech units, or any other device with similar functionality, could also be implemented.

[0037] Other aspects, modifications, and embodiments are within the scope of the following claims.

What is claimed is:

Claims

1. A method of extracting a subset of speech units from a larger set of speech units for use by a speech synthesizer in synthesizing speech, wherein the speech units are stored in a compressed encoded representation that was generated by a codec, the method comprising:

selecting members of the subset of speech units based on an overall cost associated with using the speech synthesizer to synthesize a test set of speech, wherein the overall cost includes at least one error introduced by using the codec to decode the stored representations of the speech units; and
storing the selected subset of speech units on a speech-enabled device.

2. The method of claim 1, wherein the selection step comprises:

for each member of a plurality of members of the set of speech units,
computing the overall cost associated with using the speech synthesizer to synthesize the test set of speech from a population of speech units that excludes that member; and
from the larger set of speech units, selecting the speech units that are members of the subset of speech units based upon the overall costs computed for the plurality of members.

3. The method of claim 1, wherein the at least one error introduced by using the codec comprises an error associated with decoding individual stored representations of the speech units.

4. The method of claim 1, wherein the at least one error introduced by using the codec comprises an error associated with a continuity between successive decoded stored representations of the speech units.

5. The method of claim 1, wherein the overall cost comprises a weighted sum of a target cost and a continuity cost, and the at least one error introduced by the codec includes an error associated with decoding individual stored representations of the speech units.

6. The method of claim 5, wherein the weights are selected to optimize a quality of the synthesized test set when the overall cost as defined by the selected weights is minimized.
7. The method of claim 6 wherein the quality of the synthesized test set is empirically determined by at least one listener.
8. The method of claim 6, wherein the quality of the synthesized test set is determined by an objective comparison of the synthesized test set with the original test set.
9. The method of claim 1, wherein the overall cost comprises a weighted sum of a target cost and a continuity cost, at the least one error introduced by the codec includes an error associated with a continuity between adjacent decoded speech units.
10. The method of claim 9, wherein the weights are selected to optimize a quality of the synthesized test set when the overall cost as defined by the selected weights is minimized.
11. The method of claim 10, wherein the quality of the synthesized test set is empirically determined by at least one listener.
12. The method of claim 10, wherein the quality of the synthesized test set is determined by an objective comparison of the synthesized test set with the original test set.
13. The method of claim 1, wherein each speech unit is pre-assigned to a cluster by a cluster analysis process, the assignment being based on at least one linguistic feature of the speech unit.
14. The method of claim 13, wherein the subset of speech units comprises at least one unit selected from each cluster.

15. The method of claim 14, wherein the overall cost used to perform the selection of the at least one unit from each cluster is further based at least in part on a measure of how well the unit represents the cluster acoustically.

16. The method of claim 1, wherein a number of units in the subset of speech units is determined at least in part by a storage constraint of the speech-enabled device.

1/2

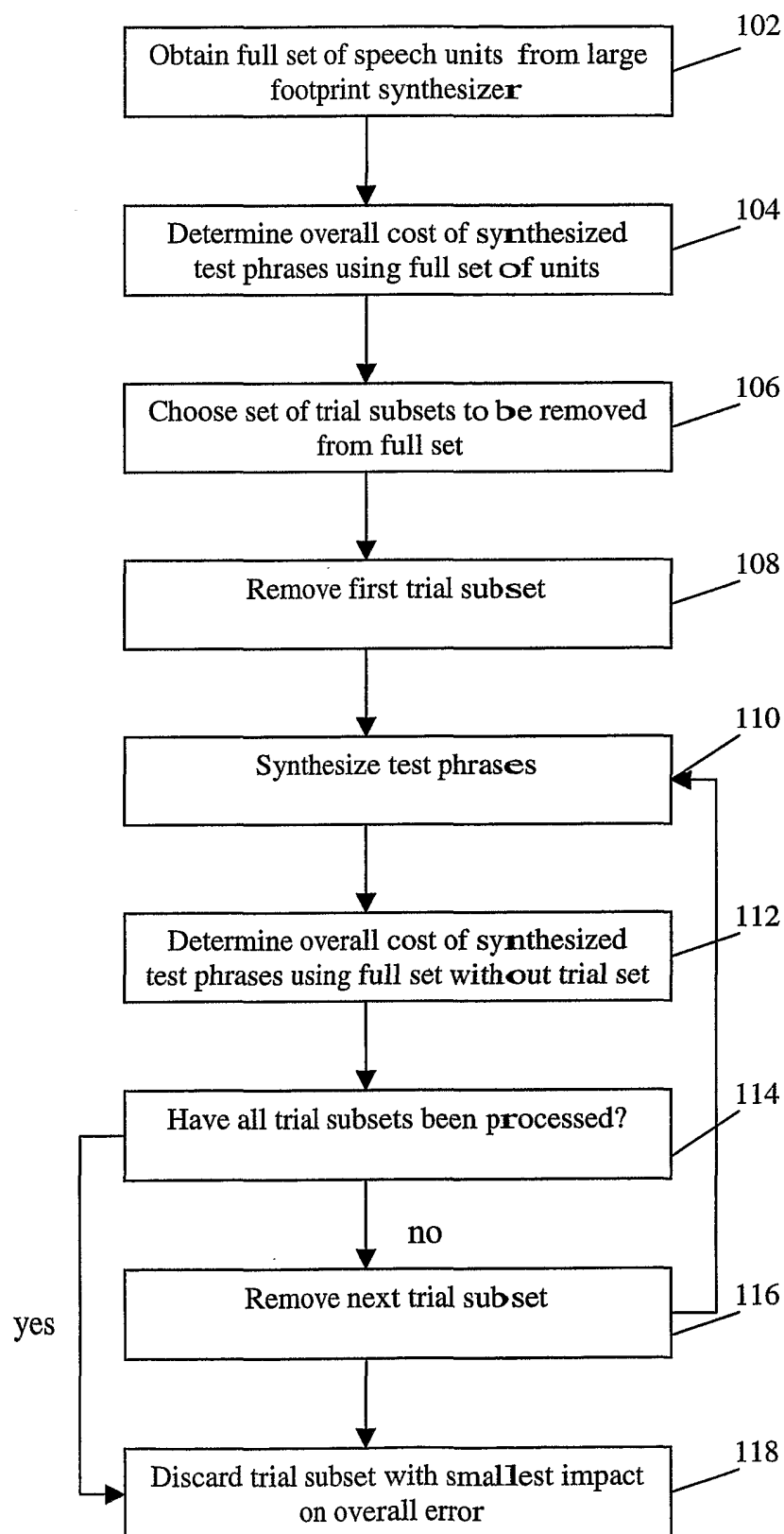


FIG. 1

2/2

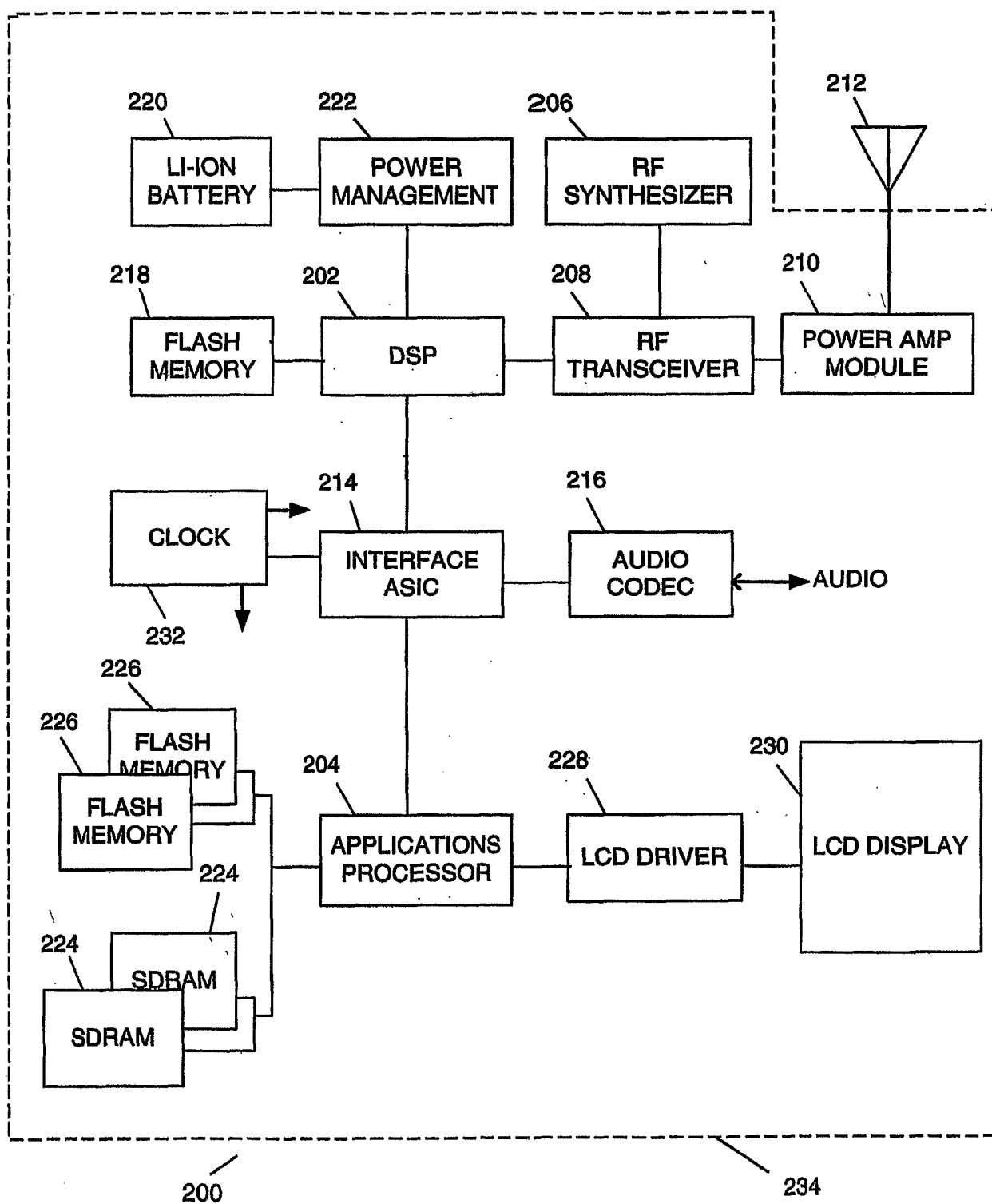


FIG. 2

INTERNATIONAL SEARCH REPORT

International application No
PCT/US2005/039207

A. CLASSIFICATION OF SUBJECT MATTER G10L13/06				
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, WPI Data, INSPEC, IBM-TDB				
C. DOCUMENTS CONSIDERED TO BE RELEVANT				
Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.		
A	BLACK A W: "Perfect synthesis for all of the people all of the time" SPEECH SYNTHESIS, 2002. PROCEEDINGS OF 2002 IEEE WORKSHOP ON 11-13 SEPT. 2002, PISCATAWAY, NJ, USA, IEEE, 11 September 2002 (2002-09-11), pages 167-170, XP010653637 ISBN: 0-7803-7395-2 abstract page 167, left-hand column, paragraph 1 - right-hand column, paragraph 4 page 170, left-hand column, paragraphs 5,6 ----- -/--	1-16		
☒ Further documents are listed in the continuation of Box C. ☐ See patent family annex.				
* Special categories of cited documents				
<table style="width: 100%; border: none;"> <tr> <td style="width: 50%; vertical-align: top; border: none;"> "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 </td> <td style="width: 50%; vertical-align: top; border: none;"> "T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention "X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone "Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art. "&" document member of the same patent family </td> </tr> </table>			"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 novel or cannot be considered to involve an inventive step when the document is taken alone "Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art. "&" document member of the same patent family
"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 novel or cannot be considered to involve an inventive step when the document is taken alone "Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art. "&" document member of the same patent family			
Date of the actual completion of the international search 28 March 2006		Date of mailing of the international search report 12/04/2006		
Name and mailing address of the ISA/ European Patent Office, P.B. 5818 Patentlaan 2 NL - 2280 HV Rijswijk Tel. (+31-70) 340-2040, Tx. 31 651 epo nl, Fax (+31-70) 340-3016		Authorized officer Greiser, N		

INTERNATIONAL SEARCH REPORT

International application No

PCT/US2005/039207

C(Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	KUMAR R, KISHORE S P: "Automatic Pruning of Unit Selection speech Databases for Synthesis without Loss of Naturalness" INTERSPEECH 2004 ICSLP, [Online] 4 October 2004 (2004-10-04), - 8 October 2004 (2004-10-08) pages 1-4, XP002374396 Jeju Island (Korea) Retrieved from the Internet: URL:http://speech.iiit.net/{rohit/work/UnitPruning-CR.pdf} [retrieved on 2006-03-20] abstract page 1, left-hand column, paragraph 1 - last paragraph page 4, right-hand column, paragraphs 3,4 -----	1
A	DONOVAN R E, ITTYCHERIAH A, FRANZ M, RAMABHADRAN B, EIDE E ET AL.: "Current status of the IBM Trainable Speech Synthesis System" PROCEEDINGS 4TH ESCA TUTORIAL AND RESEARCH WORKSHOP ON SPEECH SYNTHESIS, [Online] 2001, pages 1-4, XP002374397 Atholl Palace Hotel Scotland (GB) Retrieved from the Internet: URL:http://crow.ee.washington.edu/people/bulyko/papers/donovan-esca01a.pdf> [retrieved on 2006-03-20] abstract page 1, left-hand column, paragraph 1 page 1, right-hand column, lines 22-25 page 2, left-hand column, paragraph 7 - right-hand column, paragraph 1 -----	1