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

(11) **EP 1 860 646 A2**

(12) **EUROPEAN PATENT APPLICATION**

(43) Date of publication:
28.11.2007 Bulletin 2007/48

(51) Int Cl.:
G10L 13/06 *(2006.01)*

(21) Application number: **07116266.3**

(22) Date of filing: **27.03.2003**

(84) Designated Contracting States:
DE FI FR GB NL

(30) Priority: **29.03.2002 US 369043**
14.01.2003 US 341869

(62) Document number(s) of the earlier application(s) in accordance with Art. 76 EPC:
03100795.8 / 1 394 769

(71) Applicant: **AT&T Corp.**
New York, NY 10013-2412 (US)

(72) Inventors:
- **Conkie, Alistair, D.**
Morris County, NJ 07960 (US)
- **Kim, Yeon-Jun**
Morris County, NJ 07981 (US)

(74) Representative: **Suckling, Andrew Michael**
Marks & Clerk
4220 Nash Court
Oxford Business Park South
Oxford
Oxfordshire OX4 2RU (GB)

Remarks:
This application was filed on 12 - 09 - 2007 as a divisional application to the application mentioned under INID code 62.

(54) **Automatic segmentaion in speech synthesis**

(57) A method for segmenting phone labels to reduce misalignments in order to improve synthetic speech when the phone labels are concatenated comprises:
training a set of HMMs using one of a specific speaker's hand-labeled speech data and speaker-independent speech data;
segmenting the trained set of HMMs using an alignment to produce phone labels, wherein each phone label has a spectral boundary;
using a weighted slope metric to identify bending points of spectral transitions, wherein each bending point corresponds to a spectral boundary; and
correcting a particular spectral boundary of a particular phone label if the particular spectral boundary does not coincide with a particular bending point.

EP 1 860 646 A2

Description

Related Applications

[0001] This application claims the benefit of U.S. Provisional Patent Application Serial No. 60/369,043 entitled "System and Method of Automatic Segmentation for Text to Speech Systems" and filed March 29, 2002, which is incorporated herein by reference.

BACKGROUND OF THE INVENTION

The Field of the Invention

[0002] The present invention relates to systems and methods for automatic segmentation in speech synthesis. More particularly, the present invention relates to systems and methods for automatic segmentation in speech synthesis by combining a Hidden Markov Model (HMM) approach with spectral boundary correction.

The Relevant Technology

[0003] One of the goals of text-to-speech (TTS) systems is to produce high-quality speech using a large-scale speech corpus. TTS systems have many applications and, because of their ability to produce speech from text, can be easily updated to produce a different output by simply altering the textual input. Automated response systems, for example, often utilize TTS systems that can be updated in this manner and easily configured to produce the desired speech. TTS systems also play an integral role in many automatic speech recognition (ASR) systems.

[0004] The quality of a TTS system is often dependent on the speech inventory and on the accuracy with which the speech inventory is segmented and labeled. The speech or acoustic inventory usually stores speech units (phones, diphones, half-phones, etc.) and during speech synthesis, units are selected and concatenated to create the synthetic speech. In order to achieve high quality synthetic speech, the speech inventory should be accurately segmented and labeled in order to avoid noticeable errors in the synthetic speech.

[0005] Obtaining a well segmented and labeled speech inventory, however, is a difficult and time consuming task. Manually segmenting or labeling the units of a speech inventory cannot be performed in real time speeds and may require on the order of 200 times real time to properly segment a speech inventory. Accordingly, it will take approximately 400 hours to manually label 2 hours of speech. In addition, consistent segmentation and labeling of a speech inventory may be difficult to achieve if more than one person is working on a particular speech inventory. The ability to automate the process of segmenting and labeling speech would clearly be advantageous.

[0006] In the development of both ASR and TTS systems, automatic segmentation of a speech inventory plays an important role in significantly reducing reduce the human effort that would otherwise be require to build, train, and/or segment speech inventories. Automatic segmentation is particularly useful as the amount of speech to be processed becomes larger.

[0007] Many TTS systems utilize a Hidden Markov Model (HMM) approach to perform automatic segmentation in speech synthesis. One advantage of a HMM approach is that it provides a consistent and accurate phone labeling scheme. Consistency and accuracy are critical for building a speech inventory that produces intelligible and natural sounding speech. Consistent and accurate segmentation is particularly useful in a TTS system based on the principles of unit selection and concatenative speech synthesis.

[0008] Even though HMM approaches to automatic segmentation in speech syntheses have been successful, there is still room for improvement regarding the degree of automation and accuracy. As previously stated, there is a need to reduce the time and cost of building an inventory of speech units. This is particularly true as a demand for more synthetic voices, including customized voices, increases. This demand has been primarily satisfied by performing the necessary segmentation work manually, which significantly lengthens the time required to build the speech inventories.

[0009] For example, hand-labeled bootstrapping may require a month of labeling by a phonetic expert to prepare training data for speaker-dependent HMMs (SD HMMs). Although hand-labeled bootstrapping provides quite accurate phone segmentation results, the time required to hand label the speech inventory is substantial. In contrast, bootstrapping automatic segmentation procedures with speaker-independent HMMs (SI HMMs) instead of SD HMMs reduces the manual workload considerably while keeping the HMMs stable. Even when SI HMMs are used, there is still room for improving the segmentation accuracy and degree of segmentation automation.

[0010] Another concern with regard to automatic segmentation is that the accuracy of the automatic segmentation determines, to a large degree, the quality of speech that is synthesized by unit selection and concatenation. An HMM-based approach is somewhat limited in its ability to remove discontinuities at concatenation points because the Viterbi alignment used in an HMM-based approach tries to find the best HMM sequence when given a phone transcription and

a sequence of HMM parameters rather than the optimal boundaries between adjacent units or phones. As a result, an HMM-based automatic segmentation system may locate a phone boundary at a different position than expected, which results in mismatches at unit concatenation points and in speech discontinuities. There is therefore a need to improve automatic segmentation.

BRIEF SUMMARY OF THE INVENTION

[0011] The present invention overcomes these and other limitations and relates to systems and methods for automatically segmenting a speech inventory. More particularly, the present invention relates to systems and methods for automatically segmenting phones and more particularly to automatically segmenting a speech inventory by combining an HMM-based approach with spectral boundary correction.

[0012] In one embodiment, automatic segmentation begins by bootstrapping a set of HMMs with speaker-independent HMMs. The set of HMMs is initialized, re-estimated, and aligned to produce the labeled units or phones. The boundaries of the phone or unit labels that result from the automatic segmentation are corrected using spectral boundary correction. The resulting phones are then used as seed data for HMM initialization and re-estimation. This process is performed iteratively.

[0013] A phone boundary is defined, in one embodiment, as the position where the maximal concatenation cost concerning spectral distortion is located. Although Euclidean distance between mel frequency cepstral coefficients (MFCCs) is often used to calculate spectral distortions, the present invention utilizes a weighted slop metric. The bending point of a spectral transition often coincides with a phone boundary. The spectral-boundary-corrected phones are then used to initialize, re-estimate and align the HMMs iteratively. In other words, the labels that have been re-aligned using spectral boundary correction are used as feedback for iteratively training the HMMs. In this manner, misalignments between target phone boundaries and boundaries assigned by automatic segmentation can be reduced.

[0014] Additional features and advantages of the invention will be set forth in the description which follows, and in part will be obvious from the description, or may be learned by the practice of the invention. The features and advantages of the invention may be realized and obtained by means of the instruments and combinations particularly pointed out in the appended claims. These and other features of the present invention will become more fully apparent from the following description and appended claims, or may be learned by the practice of the invention as set forth hereinafter.

BRIEF DESCRIPTION OF THE DRAWINGS

[0015] A more particular description of the invention briefly described above will be rendered by reference to specific embodiments thereof which are illustrated in the appended drawings. Understanding that these drawings depict only typical embodiments of the invention and are not therefore to be considered limiting of its scope, the invention will be described and explained with additional specificity and detail through the use of the accompanying drawings in which:

Figure 1 illustrates a text-to-speech system that converts textual input to audible speech;
Figure 2 illustrates an exemplary method for automatic segmentation using spectral boundary correction with an HMM approach; and
Figure 3 illustrates a bending point of a spectral transition that coincides with a phone boundary in one embodiment.

DETAILED DESCRIPTION OF THE PREFERRED EMBODIMENTS

[0016] Speech inventories are used, for example, in text-to-speech (TTS) systems and in automatic speech recognition (ASR) systems. The quality of the speech that is rendered by concatenating the units of the speech inventory represents how well the units or phones are segmented. The present invention relates to systems and methods for automatically segmenting speech inventories and more particularly to automatically segmenting a speech inventory by combining an HMM-based segmentation approach with spectral boundary correction. By combining an HMM-based segmentation approach with spectral boundary correction, the segmental quality of synthetic speech in unit-concatenative speech synthesis is improved.

[0017] An exemplary HMM-based approach to automatic segmentation usually includes two phases: training the HMMs, and unit segmentation using the Viterbi alignment. Typically, each phone or unit is defined as an HMM prior to unit segmentation and then trained with a given phonetic transcription and its corresponding feature vector sequence. TTS systems often require more accuracy in segmentation and labeling than do ASR systems.

[0018] Figure 1 illustrates an exemplary TTS system that converts text to speech. In Figure 1, the TTS system 100 converts the text 110 to audible speech 118 by first performing a linguistic analysis 112 on the text 110. The linguistic analysis 112 includes, for example, applying weighted finite state transducers to the text 110. In prosodic modeling 114, each segment is associated with various characteristics such as segment duration, syllable stress, accent status, and

the like. Speech synthesis 116 generates the synthetic speech 118 by concatenating segments of natural speech from a speech inventory 120. The speech inventory 120, in one embodiment, usually includes a speech waveform and phone labeled data.

[0019] The boundary of a unit (phone, diphone, etc.) for segmentation purposes is defined as being where one unit ends and another unit begins. For the speech to be coherent and natural sounding, the segmentation must occur as close to the actual unit boundary as possible. This boundary often naturally occurs within a certain time window depending on the class of the two adjacent units. In one embodiment of the present invention, only the boundaries within these time windows are examined during spectral boundary correction in order to obtain more accurate unit boundaries. This prevents a spurious boundary from being inadvertently recognized as the phone boundary, which would lead to discontinuities in the synthetic speech.

[0020] Figure 2 illustrates an exemplary method for automatically segmenting phones or units and illustrates three examples of seed data to begin the initialization of a set of HMMs. Seed data can be obtained using, for example: hand-labeled bootstrap 202, speaker-independent (SI) HMM bootstrap 204, and a flat start 206. Hand-labeled bootstrapping, which utilizes a specific speaker's hand-labeled speech data, results in the most accurate HMM modeling and is often called speaker-dependent HMM (SD HMM). While SD HMMs are generally used for automatic segmentation in speech synthesis, they have the disadvantage of being quite time-consuming to prepare. One advantage of the present invention is to reduce the amount of time required to segment the speech inventory.

[0021] If hand-labeled speech data is available for a particular language, but not for the intended speaker, bootstrapping with SI HMM alignment is the best alternative. In one embodiment, SI HMMs for American English, trained with the TIMIT speech corpus, were used in the preparation of seed phone labels. With the resulting labels, SD HMMs for an American male speaker were trained to provide the segmentation for building an inventory of synthesis units. One advantage of bootstrapping with SI HMMs is that all of the available speech data can be used as training data if necessary.

[0022] In this example, the automatic segmentation system includes ARPA phone HMMs that use three-state left-to-right models with multiple mixture of Gaussian density. In this example, standard HMM input parameters, which include twelve MFCCs (Mel frequency cepstral coefficients), normalized energy, and their first and second order delta coefficients, are utilized.

[0023] Using one hundred randomly chosen sentences, the SD HMMs bootstrapped with SI HMMs result in phones being labeled with an accuracy of 87.3% (< 20 ms, compared to hand labeling). Many errors are caused by differences between the speaker's actual pronunciations and the given pronunciation lexicon, i.e., errors by the speaker or the lexicon or effects of spoken language such as contractions. Therefore, speaker-individual pronunciation variations have to be added to the lexicon.

[0024] Figure 2 illustrates a flow diagram for automatic segmentation that combines an HMM-based approach with iterative training and spectral boundary correction. Initialization 208 occurs using the data from the hand-labeled bootstrap 202, the SI HMM bootstrap 204, or from a flat start 206. After the HMMs are initialized, the HMMs are re-estimated (210). Next, embedded re-estimation 212 is performed. These actions - initialization 208, re-estimation 210, and embedded re-estimation 212 - are an example of how HMMs are trained from the seed data.

[0025] After the HMMs are trained, a Viterbi alignment 214 is applied to the HMMs in one embodiment to produce the phone labels 216. After the HMMs are aligned, the phones are labeled and can be used for speech synthesis. In Figure 2, however, spectral boundary correction is applied to the resulting phone labels 216. Next, the resulting phones are trained and aligned iteratively. In other words, the phone labels that have been re-aligned using spectral boundary correction are used as input to initialization 208 iteratively. The hand-labeled bootstrapping 202, SI HMM bootstrapping 204, and the flat start 206 are usually used the first time the HMMs are trained. Successive iterations use the phone labels that have been aligned using spectral boundary correction 218.

[0026] The motivation for iterative HMM training is that more accurate initial estimates of the HMM parameters produce more accurate segmentation results. The phone labels that result from bootstrapping with SI HMMs are more accurate than the original input (seed phone labels). For this reason, for tuning the SD HMMs to produce the best results, the phone labels resulting from the previous iteration and corrected using spectral boundary correction 218 are used as the input for HMM initialization 208 and re-estimation 210, as shown in Figure 2. This procedure is iterated to fine-tune the SD HMMs in this example.

[0027] After several rounds of iterative training that includes spectral boundary correction, mismatches between manual labels and phone labels assigned by an HMM-based approach will be considerably reduced. For example, when the HMM training procedure illustrated in Figure 2 was iterated five times in one example, an accuracy of 93.1 % was achieved, yielding a noticeable improvement in synthesis quality. The accuracy of phone labeling in a few speech samples alone cannot predict synthetic quality itself. The stop condition for iterative training, therefore, is defined as the point when no more perceptual improvement of synthesis quality can be observed.

[0028] A reduction of mismatches between phone boundary labels is expected when the temporal alignment of the feed-back labeling is corrected. Phone boundary corrections can be done manually or by rule-based approaches. Assuming that the phone labels assigned by an HMM-based approach are relatively accurate, automatic phone boundary

correction concerning spectral features improves the accuracy of the automatic segmentation.

[0029] One advantage of the present invention is to reduce or minimize the audible signal discontinuities caused by spectral mismatches between two successive concatenated units. In unit-concatenative speech synthesis, a phone boundary can be defined as the position where the maximal concatenation cost concerning spectral distortion, i.e., the spectral boundary, is located. The Euclidean distance between MFCCs is most widely used to calculate spectral distortions. As MFCCs were likely used in the HMM-based segmentation, the present embodiment uses instead the weighted slope metric (see Equation (1) below).

$$\mathrm{d}(S^L, S^R) = u_E \left| E_{S^L} - E_{S^R} \right| + \sum_{i=1}^{K} u(i) \left[\Delta_{S^L}(i) - \Delta_{S^R}(i) \right]^2 \qquad (1)$$

[0030] In this example, S^L and S^R are 256 point FFTs (fast Fourier transforms) divided into K critical bands. The S^L and S^R vectors represent the spectrum to the left and the right of the boundary, respectively. E_{SL} and E_{SR} are spectral energy, $\Delta_{SL}(i)$ and $\Delta_{SR}(i)$ are the ith critical band spectral slopes of S^L and S^R (see Figure 3), and u_E, $u(i)$ are weighting factors for the spectral energy difference and the ith spectral transition.

[0031] Spectral transitions play an important role in human speech perception. The bending point of spectral transition, i.e., the local maximum of

$$\sum_{i=1}^{K} u(i) \left[\Delta_{S^L}(i) - \Delta_{S^R}(i) \right]^2 ,$$

often coincides with a phone boundary. Figure 3, which illustrates adjacent spectral slopes, more fully illustrates the bending point of a spectral transition. In this example, the spectral slope 304 corresponds to the ith critical band of S^L, and the spectral slope 306 corresponds to the ith critical band of S^R. The bending point 302 of the spectral transition usually coincides with a phone boundary. Using spectral boundaries identified in this fashion, spectral boundary correction 218 can be applied to the phone labels 216, as illustrated in Figure 2.

[0032] In the present embodiment, $|E_{SL} - E_{SR}|$, which is the absolute energy difference in Equation (1), is modified to distinguish K critical bands, as in Equation (2):

$$\left| E_{S^L} - E_{S^R} \right| = \sum_{j=1}^{K} w(j) * \left| E_{S^L}(j) - E_{S^R}(j) \right| \qquad (2)$$

where $w(j)$ is the weight of the jth critical band. This is because each phone boundary is characterized by energy changes in different bands of the spectrum.

[0033] Although there is a strong tendency for the largest peak to occur at the correct phone boundary, the automatic detector described above may produce a number of spurious peaks. To minimize the mistakes in the automatic spectral boundary correction, a context-dependent time window in which the optimal phone boundary is more likely to be found is used. The phone boundary is checked only within the specified context-dependent time window.

[0034] Temporal misalignment tends to vary in time depending on the contexts of two adjacent phones. Therefore, the time window for finding the local maximum of spectral boundary distortion is empirically determined, in this embodiment, by the adjacent phones as illustrated in the following table. This table represents context-dependent time windows (in ms) for spectral boundary correction (V: Vowel, P: Unvoiced stop, B: Voiced stop, S: Unvoiced fricative, Z: Voiced fricative, L: Liquid, N: Nasal).

BOUNDARY	Time window (ms)	BOUNDARY	Time window (ms)
V-V	-4.5 ± 50	P-V	-1.6 ± 30
V-N	-4.8 ± 30	N-V	0 ± 30
V-B	-13.9 ± 30	B-V	0 ± 20

(continued)

BOUNDARY	Time window (ms)	BOUNDARY	Time window (ms)
V-L	-23.2 ± 40	L-V	11.1 ± 30
V-P	2.2 ± 20	S-V	2.7 ± 20
V-Z	-15.8 ± 30	Z-V	15.4 ± 40

[0035] The present invention relates to a method for automatically segmenting phones or other units by combining HMM-based segmentation with spectral features using spectral boundary correction. Misalignments between target phone boundaries and boundaries assigned by automatic segmentation are reduced and result in more natural synthetic speech. In other words, the concatenation points are less noticeable and the quality of the synthetic speech is improved.

[0036] The embodiments of the present invention may comprise a special purpose or general purpose computer including various computer hardware, as discussed in greater detail below. Embodiments within the scope of the present invention may also include computer-readable media for carrying or having computer-executable instructions or data structures stored thereon. Such computer-readable media can be any available media that can be accessed by a general purpose or special purpose computer. By way of example, and not limitation, such computer-readable media can comprise RAM, ROM, EEPROM, CD-ROM or other optical disk storage, magnetic disk storage or other magnetic storage devices, or any other medium which can be used to carry or store desired program code means in the form of computer-executable instructions or data structures and which can be accessed by a general purpose or special purpose computer. When information is transferred or provided over a network or another communications connection (either hardwired, wireless, or a combination of hardwired or wireless) to a computer, the computer properly views the connection as a computer-readable medium. Thus, any such connection is properly termed a computer-readable medium. Combinations of the above should also be included within the scope of computer-readable media.

[0037] Computer-executable instructions include, for example, instructions and data which cause a general purpose computer, special purpose computer, or special purpose processing device to perform a certain function or group of functions. Computer-executable instructions also include program modules which are executed by computers in stand alone or network environments. Generally, program modules include routines, programs, objects, components, data structures, etc. that perform particular tasks or implement particular abstract data types. Computer-executable instructions, associated data structures, and program modules represent examples of the program code means for executing steps of the methods disclosed herein. The particular sequence of such executable instructions or associated data structures represents examples of corresponding acts for implementing the functions described in such steps.

[0038] The present invention may be embodied in other specific forms without departing from its spirit or essential characteristics. The described embodiments are to be considered in all respects only as illustrative and not restrictive. The scope of the invention is, therefore, indicated by the appended claims rather than by the foregoing description. All changes which come within the meaning and range of equivalency of the claims are to be embraced within their scope.

Claims

1. A method for segmenting phone labels to reduce misalignments in order to improve synthetic speech when the phone labels are concatenated, the method comprising:

   training a set of HMMs using one of a specific speaker's hand-labeled speech data and speaker-independent speech data;
   segmenting the trained set of HMMs using an alignment to produce phone labels, wherein each phone label has a spectral boundary;
   using a weighted slope metric to identify bending points of spectral transitions, wherein each bending point corresponds to a spectral boundary; and
   correcting a particular spectral boundary of a particular phone label if the particular spectral boundary does not coincide with a particular bending point.

2. A method as defined in claim 1, wherein using a weighted slope metric to identify bending points of spectral transitions further comprises applying the weighted slope metric within context-dependent time windows such that spurious spectral boundaries are not applied to the phone labels.

3. A method as defined in claim 2, further comprising retraining the set of HMMs using the phone labels that have

been corrected using the weighted slope metric.

4. A method as defined in claim 2, wherein each spectral boundary is defined by a first phone class and a second phone class, wherein the first phone class and the second phone class include at least one of a vowel, an unvoiced stop, a voiced stop, an unvoiced fricative, a voiced fricative, a liquid class and a nasal class.

5. A method as defined in claim 2, further comprising determining context-dependent time windows empirically.

6. A computer-readable media having computer-executable instructions for performing the method of claim 1.

7. A system for segmenting phone labels to reduce misalignments in order to improve synthetic speech when the phone labels are concatenated, the system comprising:

means for training a set of HMMs using one of a specific speaker's hand-labeled speech data and speaker-independent speech data;
means for segmenting the trained set of HMMs using an alignment to produce phone labels, wherein each phone label has a spectral boundary;
means for using a weighted slope metric to identify bending points of spectral transitions, wherein each bending point corresponds to a spectral boundary; and
means for correcting a particular spectral boundary of a particular phone label if the particular spectral boundary does not coincide with a particular bending point.

8. A system as defined in claim 7, wherein the means for using a weighted slope metric to identify bending points of spectral transitions are adapted to apply the weighted slope metric within context-dependent time windows such that spurious spectral boundaries are not applied to the phone labels.

9. A system as defined in claim 8, further comprising means for retraining the set of HMMs using the phone labels that have been corrected using the weighted slope metric.

10. A system as defined in claim 8, wherein each spectral boundary is defined by a first phone class and a second phone class, wherein the first phone class and the second phone class include at least one of a vowel, an unvoiced stop, a voiced stop, an unvoiced fricative, a voiced fricative, a liquid class and a nasal class.

11. A system as defined in claim 8, further comprising means for determining context-dependent time windows empirically.

FIG. 1

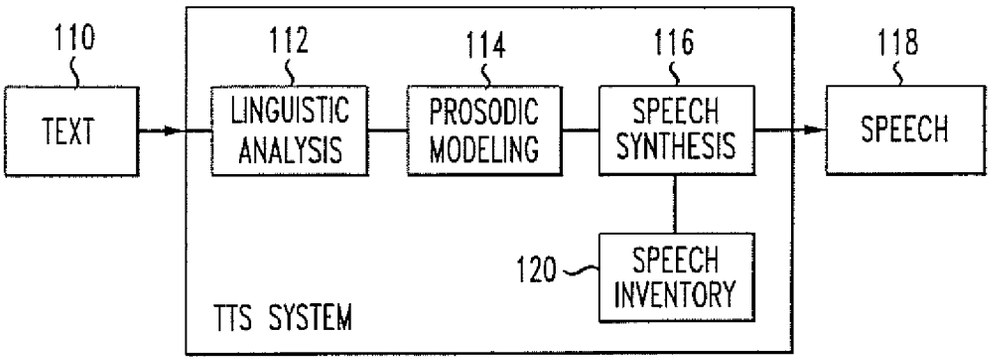

110
TEXT
112
LINGUISTIC ANALYSIS
114
PROSODIC MODELING
116
SPEECH SYNTHESIS
118
SPEECH
120
SPEECH INVENTORY
TTS SYSTEM

FIG. 2

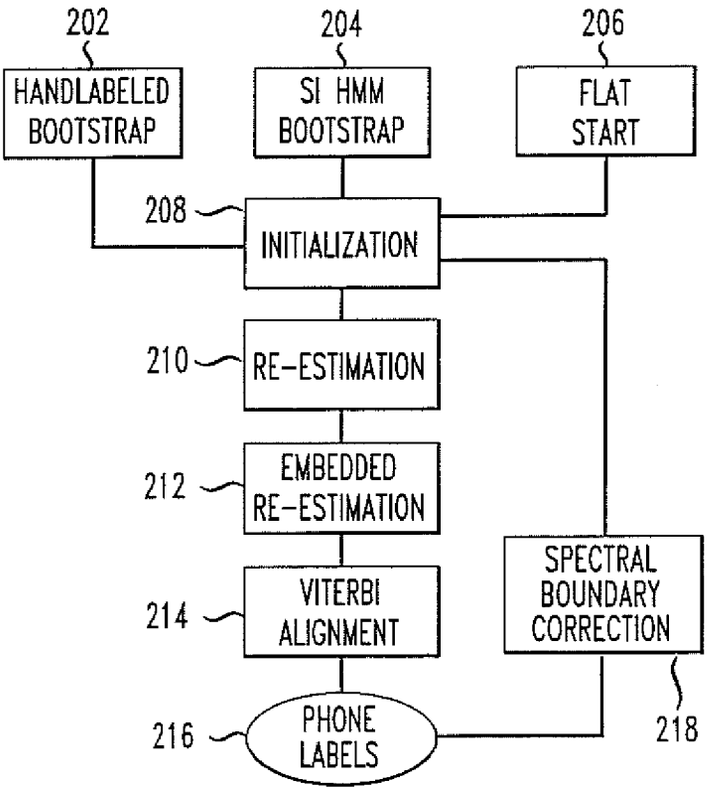
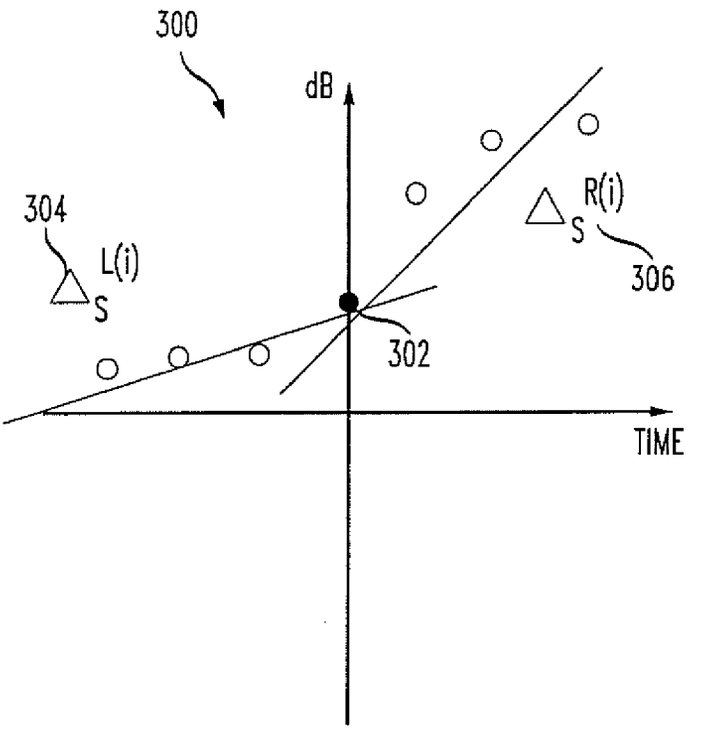

202
HANDLABELED BOOTSTRAP
204
SI HMM BOOTSTRAP
206
FLAT START
208
INITIALIZATION
210
RE-ESTIMATION
212
EMBEDDED RE-ESTIMATION
214
VITERBI ALIGNMENT
216
PHONE LABELS
SPECTRAL BOUNDARY CORRECTION
218

FIG. 3



REFERENCES CITED IN THE DESCRIPTION

This list of references cited by the applicant is for the reader's convenience only. It does not form part of the European patent document. Even though great care has been taken in compiling the references, errors or omissions cannot be excluded and the EPO disclaims all liability in this regard.

Patent documents cited in the description

- US 369043 P **[0001]**