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

(11) EP 3 149 727 B1

(12) **EUROPEAN PATENT SPECIFICATION**

(45) Date of publication and mention of the grant of the patent:
27.01.2021 Bulletin 2021/04

(21) Application number: **14893138.9**

(22) Date of filing: **28.05.2014**

(51) Int Cl.:
G10L 13/02 *(2013.01)* ***G10L 25/90*** *(2013.01)*

(86) International application number:
PCT/US2014/039722

(87) International publication number:
WO 2015/183254 (03.12.2015 Gazette 2015/48)

(54) **METHOD FOR FORMING THE EXCITATION SIGNAL FOR A GLOTTAL PULSE MODEL BASED PARAMETRIC SPEECH SYNTHESIS SYSTEM**

VERFAHREN ZUR ERZEUGUNG DES ANREGUNGSSIGNALS FÜR EIN GLOTTALES IMPULSMODELLBASIERTES PARAMETRISCHES SPRACHSYNTHESESYSTEM

PROCÉDÉ PERMETTANT DE FORMER UN SIGNAL D'EXCITATION DESTINÉ À UN SYSTÈME DE SYNTHÈSE VOCALE PARAMÉTRIQUE BASÉ SUR UN MODÈLE D'IMPULSION GLOTTALE

(84) Designated Contracting States:
AL AT BE BG CH CY CZ DE DK EE ES FI FR GB GR HR HU IE IS IT LI LT LU LV MC MK MT NL NO PL PT RO RS SE SI SK SM TR

(43) Date of publication of application:
05.04.2017 Bulletin 2017/14

(73) Proprietor: **Interactive Intelligence Group, Inc.**
Indianapolis, IN 46278 (US)

(72) Inventors:
- **DACHIRAJU, Rajesh**
  Hyderabad 500038 (IN)
- **GANAPATHIRAJU, Aravind**
  Hyderabad
  AP (IN)

(74) Representative: **FRKelly**
27 Clyde Road
Dublin D04 F838 (IE)

(56) References cited:
US-A- 5 400 434
US-A1- 2012 123 782
US-A1- 2014 142 946
US-B1- 6 795 807
US-B2- 8 386 256

- **RAITIO TUOMO ET AL: "Comparing glottal-flow-excited statistical parametric speech synthesis methods", 2013 IEEE INTERNATIONAL CONFERENCE ON ACOUSTICS, SPEECH AND SIGNAL PROCESSING (ICASSP); VANCOUCER, BC; 26-31 MAY 2013, INSTITUTE OF ELECTRICAL AND ELECTRONICS ENGINEERS, PISCATAWAY, NJ, US, 26 May 2013 (2013-05-26), pages 7830-7834, XP032509201, ISSN: 1520-6149, DOI: 10.1109/ICASSP.2013.6639188 [retrieved on 2013-10-18] & TUOMO RAITIO ET AL: "Utilizing glottal source pulse library for generating improved excitation signal for HMM-based speech synthesis", 2011 IEEE INTERNATIONAL CONFERENCE ON ACOUSTICS, SPEECH AND SIGNAL PROCESSING : (ICASSP 2011) ; PRAGUE, CZECH REPUBLIC, 22 - 27 MAY 2011, IEEE, PISCATAWAY, NJ, 22 May 2011 (2011-05-22), pages 4564-4567, XP032001695, DOI: 10.1109/ICASSP.2011.5947370 ISBN: 978-1-4577-0538-0**
- **TAMAS GABOR CSAPO ET AL: "A novel codebook-based excitation model for use in speech synthesis", COGNITIVE INFOCOMMUNICATIONS (COGINFOCOM), 2012 IEEE 3RD INTERNATIONAL CONFERENCE ON, IEEE, 2 December 2012 (2012-12-02), pages 661-665, XP032316820, DOI: 10.1109/COGINFOCOM.2012.6421934 ISBN: 978-1-4673-5187-4**

Printed by Jouve, 75001 PARIS (FR)

(Cont. next page)

- **CABRAL JOAO P ET AL: "Glottal Spectral Separation for Speech Synthesis", IEEE JOURNAL OF SELECTED TOPICS IN SIGNAL PROCESSING, IEEE, US, vol. 8, no. 2, 1 April 2014 (2014-04-01), pages 195-208, XP011542567, ISSN: 1932-4553, DOI: 10.1109/JSTSP.2014.2307274 [retrieved on 2014-03-11]**
- **PRATHOSH A P ET AL: "Epoch Extraction Based on Integrated Linear Prediction Residual Using Plosion Index", IEEE TRANSACTIONS ON AUDIO, SPEECH AND LANGUAGE PROCESSING, IEEE, vol. 21, no. 12, 1 December 2013 (2013-12-01), pages 2471-2480, XP011531024, ISSN: 1558-7916, DOI: 10.1109/TASL.2013.2273717 [retrieved on 2013-10-23]**

Description

BACKGROUND

[0001] The present invention generally relates to telecommunications systems and methods, as well as speech synthesis. More particularly, the present invention pertains to the formation of the excitation signal in a Hidden Markov Model based statistical parametric speech synthesis system.

[0002] In the prior art, Tuomo Raitio et al. present a study of the performance of glottal flow signal based excitation methods in statistical parametric speech synthesis in their contribution "Comparing Glottal-Flow-Excited Statistical Parametric Speech Synthesis Methods", 2103 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Vancouver, BC, 26-31 May 2013, Institute of Electrical and Electronics Engineers, Piscataway, NJ, US, 26 May 2013, pages 7830-7834. Further, Tamas Gabor Csapo and Geza Nemeth disclose an excitation model for the use in speech synthesis in "A Novel Codebook-Based Excitation Model for us in Speech Synthesis", Cognitive Info Communications (COGINFOCOM), 2012 IEEE 3rd International Conference on, IEEE, 2 December 2012, pages 661 to 665. None of those contributions to the prior art discloses a method to form parametric models including a combination of steps of clustering a glottal pulse database and forming a corresponding vector database.

SUMMARY

[0003] A method is presented to form parametric models, comprising the steps of: computing a glottal pulse distance metric between a number of glottal pulses; clustering a glottal pulse database into a number of clusters to determine centroid glottal pulses; forming a corresponding vector database by associating a vector with each glottal pulse in the glottal pulse database, wherein the vector associated with each glottal pulse is defined based on the glottal pulse, the centroid glottal pulses and the distance metric; and forming parametric models by associating a glottal pulse from the glottal pulse database to each determined Eigenvector.

BRIEF DESCRIPTION OF THE DRAWINGS

[0004]

Figure 1 is a diagram illustrating an embodiment of an Hidden Markov Model based Text to Speech system.
Figure 2 is a diagram illustrating an embodiment of a signal.
Figure 3 is a diagram illustrating an embodiment of excitation signal creation.
Figure 4 is a diagram illustrating an embodiment of excitation signal creation.
Figure 5 is a diagram illustrating an embodiment of overlap boundaries.
Figure 6 is a diagram illustrating an embodiment of excitation signal creation.
Figure 7 is a diagram illustrating an embodiment of glottal pulse identification.
Figure 8 is a diagram illustrating an embodiment of glottal pulse database creation.

DETAILED DESCRIPTION

[0005] For the purposes of promoting an understanding of the principles of the invention, reference will now be made to the embodiment illustrated in the drawings and specific language will be used to describe the same. It will nevertheless be understood that no limitation of the scope of the invention is thereby intended. Any alterations and further modifications in the described embodiments, and any further applications of the principles of the invention as described herein are contemplated as would normally occur to one skilled in the art to which the invention relates.

[0006] Excitation is generally assumed to be a quasi-periodic sequence of impulses for voiced regions. Each sequence is separated from the previous sequence by some duration, such as $T_0 = \frac{1}{F_0}$, where T_0 represents pitch period and F_0 represents fundamental frequency. The excitation, in unvoiced regions, is modeled as white noise. In voiced regions, the excitation is not actually impulse sequences. The excitation is instead a sequence of voice source pulses which occur due to vibration of the vocal folds. The pulses' shapes may vary depending on various factors such as the speaker, the mood of the speaker, the linguistic context, emotions, etc.

[0007] Source pulses have been treated mathematically as vectors by length normalization (through resampling) and impulse alignment, as described in European Patent EP 2242045 (granted June 27, 2012, inventors Thomas Drugman, et al.) The final length of normalized source pulse signal is resampled to meet the target pitch. The source pulse is not chosen from a database, but obtained over a series of calculations which compromise the pulse characteristics in the

frequency domain. In addition, the approximate excitation signal used for creating a pulse database does not capture low frequency source content as there is no pre-filtering done while determining the Linear Prediction (LP) coefficients, which are used for inverse filtering.

[0008] In statistical parametric speech synthesis, speech unit signals are represented by a set of parameters which can be used to synthesize speech. The parameters may be learned by statistical models, such as HMMs, for example. In an embodiment, speech may be represented as a source-filter model, wherein source/excitation is a signal which when passed through an appropriate filter produces a given sound. Figure 1 is a diagram illustrating an embodiment of a Hidden Markov Model (HMM) based Text to Speech (TTS) system. An embodiment of an exemplary system may contain two phases, for example, the training phase and the synthesis phase.

[0009] The Speech Database 105 may contain an amount of speech data for use in speech synthesis. During the training phase, a speech signal 106 is converted into parameters. The parameters may be comprised of excitation parameters and spectral parameters. Excitation Parameter Extraction 110 and Spectral Parameter Extraction 115 occurs from the speech signal 106 which travels from the Speech Database 105. A Hidden Markov Model 120 may be trained using these extracted parameters and the Labels 107 from the Speech Database 105. Any number of HMM models may result from the training and these context dependent HMMs are stored in a database 125.

[0010] The synthesis phase begins as the context dependent HMMs 125 are used to generate parameters 140. The parameter generation 140 may utilize input from a corpus of text 130 from which speech is to be synthesized from. The text 130 may undergo analysis 135 and the extracted labels 136 are used in the generation of parameters 140. In one embodiment, excitation and spectral parameters may be generated in 140.

[0011] The excitation parameters may be used to generate the excitation signal 145, which is input, along with the spectral parameters, into a synthesis filter 150. Filter parameters are generally Mel frequency cepstral coefficients (MFCC) and are often modeled by a statistical time series by using HMMs. The predicted values of the filter and the fundamental frequency as time series values may be used to synthesize the filter by creating an excitation signal from the fundamental frequency values and the MFCC values used to form the filter.

[0012] Synthesized speech 155 is produced when the excitation signal passes through the filter. The formation of the excitation signal 145 is integral to the quality of the output, or synthesized, speech 155. Low frequency information of the excitation is not captured. It will thus be appreciated that an approach is needed to capture the low frequency source content of the excitation signal and to improve the quality of synthetic speech.

[0013] Figure 2 is a graphical illustration of an embodiment of the signal regions of a speech segment, indicated generally at 200. The signal has been broken down into segments based on fundamental frequency values for categories such as voiced, unvoiced, and pause segments. The vertical axis 205 illustrates fundamental frequency in Hertz (*Hz*) while the horizontal axis 210 represents the passage of milliseconds (*ms*). The time series, F_0, 215 represents the fundamental frequency. The voiced region, 220 can be seen as a series of peaks and may be referred to as a non-zero segment. The non-zero segments 220 may be concatenated to form an excitation signal for the entire speech, as described in further detail below. The unvoiced region 225 is seen as having no peaks in the graphical illustration 200 and may be referred to as zero segments. The zero segments may represent a pause or an unvoiced segment given by the phone labels.

[0014] Figure 3 is a diagram illustrating an embodiment of excitation signal creation indicated generally at 300. Figure 3 illustrates the creation of the excitation signal for both unvoiced and pause segments. The fundamental frequency time series values, represented as F_0, represent signal regions 305 that are broken down into voiced, unvoiced, and pause segments based on the F_0 values.

[0015] An excitation signal 320 is created for unvoiced and pause segments. Where pauses occur, zeroes (0) are placed in the excitation signal. In unvoiced regions, white noise of appropriate energy (in one embodiment, this may be determined empirically by listening tests) is used as the excitation signal.

[0016] The signal regions, 305, along with the Glottal Pulse 310 are used for excitation generation 315 and subsequent generation of the excitation signal 320. The Glottal Pulse 310 comprises an Eigen glottal pulse that has been identified from the glottal pulse database, the creation of which is described in further detail in Figure 8 below.

[0017] Figure 4 is a diagram illustrating an embodiment of excitation signal creation for a voiced segment, indicated generally at 400. It is assumed that a Eigen glottal pulse has been identified from the glottal pulse database (described in further detail in Figure 7 below). The signal region 405 comprises F_0 values, which may be predicted by models, from the voiced segment. The lengths of the F_0 segments, which may be represented by N_f, are used to determine the length of the excitation signal using the mathematical equation:

$$F_0(n) = f_s * N_f * 5/1000.$$

[0018] Where f_s represents the sampling frequency of the signal. In a non-limiting example, the value of 5/1000

represents the interval of 5 *ms* durations that the F_0 values are determined for. It should be noted that any interval of a designated duration of a unit time may be used. Another array, designated as $F_0'(n)$, is obtained by linearly interpolating the F_0 array.

[0019] From the F_0 values, glottal boundaries are created, 410, which mark the pitch boundaries of the excitation signal of the voiced segments in the signal region 405. The pitch period array may be computed using the following mathematical equation:

$$T_0(n) = \frac{f_s}{F_0'(n)}$$

[0020] Pitch boundaries may then be computed using the determined pitch period array as follows:

$$P^0(i) = \sum_{j=0}^{i} T_0(P^0(i-1)$$

[0021] Where $P^0(0) = 1$, $i = 1,2,3, \ldots K$, and where P(K+1) just crosses length of the array $T_0(n)$.

[0022] The glottal pulse 415 is used along with the identified glottal boundaries 410 in the overlap adding 420 of a glottal pulse beginning at each glottal boundary. The excitation signal 425 is then created through the process of "stitching", or segment joining, to avoid boundary effects which are further described in Figures 5 and 6.

[0023] Figure 5 is a diagram illustrating an embodiment of overlap boundaries, indicated generally at 500. The illustration 500 represents a series of glottal pulses 515 and overlapping glottal pulses 520 in the segment. The vertical axis 505 represents the amplitude of excitation. The horizontal axis 510 may represent the frame number.

[0024] Figure 6 is a diagram illustrating an embodiment of excitation signal creation for a voiced segment, indicated generally at 600. "Stitching" may be used to form the final excitation signal of voiced segments (from Figure 4), which is ideally devoid of boundary effects. In an embodiment, any number of different excitation signals may have been formed through the overlap add method illustrated in Figure 4 and in the diagram 500 (Figure 5). The different excitation signals may have a constantly increasing amount of shifts in glottal boundaries 605 and an equal amount of circular left shift 630 for the glottal pulse signal. In one embodiment, if the glottal pulse signal 615 is of a length less than the corresponding pitch period, then the glottal pulse may be zero extended 625 to the length of the pitch period before circular left shifting 630 is performed. Different arrays of pitch boundaries (represented as $P^m(i)$, $m = 1,2, \ldots M - 1$) are formed with each of the same length as P^0. The arrays are computed using the following mathematical equation:

$$P^m(i) = P^0(i) + m * w$$

[0025] Where w is generally taken as 1 *msec* or, in terms of samples, $\frac{f_s}{1000}$. For a sampling frequency of $f_s = 16{,}000$, $w = 16$, for example. The highest pitch period present in the given voice segment is represented as $m * w$. Glottal pulses are created and associated with each pitch boundary array P^m. The glottal pulses 620 may be obtained from the glottal pulse signal of some length N by first zero extending it to the pitch period and then circularly left shifting it by $m * w$ samples.

[0026] For each set of frame boundaries, an excitation signal 635 is formed by initializing the glottal pulses to zero (0). Overlap add 610 is used to add the glottal pulse 620 to the first N samples of the excitation, starting from each pitch boundary value of the array $P^m(i)$, $i = 1,2,..K$. The formed signal is as a single stitched excitation, corresponding to the shift, *m*.

[0027] In an embodiment, the arithmetic mean of all of the single stitched excitation signals is then computed 640, which represents the final excitation signal for the voiced segment 645.

[0028] Figure 7 is a diagram illustrating an embodiment of glottal pulse identification, indicated generally at 700. In an embodiment, any two given glottal pulses may be used to compute the distance metric/dissimilarity between them. These are taken from the glottal pulse database 840 created in process 800 (further described in Figure 8 below). The computation may be performed by decomposing the two given glottal pulses x_i, y_i into sub-band components $x_i^{(1)}$, $x_i^{(2)}$, $x_i^{(3)}$ and $y_i^{(1)}$, $y_i^{(2)}$, $y_i^{(3)}$. The given glottal pulse may be transformed into the frequency domain by using a method such as Discrete Cosine Transform (DCT), for example. The frequency band may be split into a number of bands, which are demodulated and converted into time domain. In this example, three bands are used for illustrative purposes.

[0029] The sub-band distance metric is then computed between corresponding sub-band components of each glottal

pulses, denoted as $d_s(x_i^{(1)}, y_i^{(1)})$. The sub-band metric, which may be represented as $d_s(f,g)$, where d_s represents the distance between the two sub-band components *f* and *g*, may be computed as described in the following paragraphs.

[0030] The normalized circular cross correlation function between *f* and *g* is computed. In one embodiment, this may be denoted as $R_{f,g}(n) = f * g$, where '*' denotes normalized circular cross correlation operation between two signals. The period for circular cross correlation is taken to be the highest of lengths of the two signals *f* and *g*. The shorter signal is zero extended. The Discrete Hilbert Transform of normalized circular cross correlation is computed and denoted as $R_{f,g}^h(n)$. Using the normalized circular cross correlation and the Discrete Hilbert Transform of the normalized circular cross correlation, the signal may be determined as:

$$H_{f,g}(n) = \sqrt{R_{f,g}(n)^2 + R_{f,g}^h(n)^2}.$$

[0031] The cosine of the angle between the two signals *f* and *g* may be determined using the mathematical equation:

$$cos\theta(f,g) = maximum\ value\ of\ the\ signal\ H_{f,g}(n)\ over\ all\ n.$$

[0032] The sub-band metric, $d_s(f,g)$, between the two sub-band components *f* and *g* may be determined as:

$$d_s(f,g) = \sqrt{2(1-\cos\theta(f,g)}.$$

[0033] The distance metric between the glottal pulses is finally determined mathematically as:

$$d(x_i, y_i) = \sqrt{d_s^2\left(x_i^{(1)}, y_i^{(1)}\right) + d_s^2\left(x_i^{(2)}, y_i^{(2)}\right) + d_s^2\left(x_i^{(3)}, y_i^{(3)}\right)}$$

[0034] The glottal pulse database 840 may be clustered into a number of clusters, for example 256 (or M), using a modified k-means algorithm 705. Instead of using the Euclidean distance metric, the distance metric defined above is used. The centroids of a cluster are then updated with that element of the cluster whose sum of squares of distances from all other elements of that cluster is minimum such that: $D_m = \sum_{i=1}^{N} d^2(g_i, g_m)$ is minimum for m = c, the cluster centroid.

[0035] In an embodiment, the clustering iterations are terminated when there is no shift in any of the centroids of the *k* clusters.

[0036] A vector, a set of N real numbers, for example 256, is associated with every glottal pulse 710 in the glottal pulse database 840 to form a corresponding vector database 715. In one embodiment, the associating is performed for a given glottal pulse x_i, a vector
$V_i = [\Psi_1(x_i), \Psi_2(x_i), \Psi_3(x_i), ... \Psi_j(x_i), ... \Psi_{256}(x_i)]$, where $\Psi_j(x_i) = d^2(x_i,c_j) - d^2(x_i,x_0) - d^2(c_j,x_0)$ and, x_0 is a fixed glottal pulse picked from the database and $d^2(x_i,c_j)$ represents the square of the distance metric defined above between two glottal pulses x_i and c_j and assuming that $c_1, c_2, ... c_i, .. c_{256}$ are the centroid glottal pulses determined by clustering.

[0037] Thus, the vector associated with the given glottal pulse x_i may be computed with the mathematical equation:

$$V_i = \left[\Psi_1(x_i), \Psi_2(x_i), \Psi_3(x_i), ... \Psi_j(x_i), ... \Psi_{256}(x_i)\right]$$

[0038] In step 720, Principal Component Analysis (PCA) is performed to compute Eigenvectors of the vector database 715. In one embodiment, any one Eigenvector may be chosen 725. The closest matching vector 730 to the chosen Eigenvector from the vector database 715 is then determined in the sense of Euclidean distance. The glottal pulse from the pulse database 840 which corresponds to the closest matching vector 730 is regarded as the resulting Eigen glottal pulse 735 associated with an Eigenvector.

[0039] Figure 8 is a diagram illustrating an embodiment of glottal pulse database creation indicated generally at 800. A speech signal, 805, undergoes pre-filtering, such as pre-emphasis 810. Linear Prediction (LP) Analysis, 815, is performed using the pre-filtered signal to obtain the LP coefficients. Thus, low frequency information of the excitation may

be captured. Once the coefficients are determined, they are used to inverse filter, 820, the original speech signal, 805, which is not pre-filtered, to compute the Integrated Linear Prediction Residual (ILPR) signal 825. The ILPR signal 825 may be used as an approximation to the excitation signal, or voice source signal. The ILPR signal 825 is segmented 835 into glottal pulses using the glottal segment/cycle boundaries that have been determined from the speech signal 805. The segmentation 835 may be performed using the Zero Frequency Filtering Technique (ZFF) technique. The resulting glottal pulses may then be energy normalized. All of the glottal pulses for the entire speech training data are combined in order to form the glottal pulse database 840.

[0040] While the invention has been illustrated and described in detail in the drawings and foregoing description, the same is to be considered as illustrative and not restrictive in character, it being understood that only the preferred embodiment has been shown and described and that all equivalents, changes, and modifications that come by the following claims are desired to be protected.

[0041] Hence, the proper scope of the present invention should be determined only by the broadest interpretation of the appended claims so as to encompass all such modifications as well as all relationships equivalent to those illustrated in the drawings and described in the specification.

Claims

1. A method (700) to form parametric models, comprising the steps of:

a. computing a glottal pulse distance metric between a number of glottal pulses (710);
b. clustering a glottal pulse database (840) comprising the glottal pulses (710) into a number of clusters to determine centroid glottal pulses;
c. forming a corresponding vector database (715) by associating a vector with each glottal pulse (710) in the glottal pulse database (840), wherein the vector associated with each glottal pulse (710) is defined based on the glottal pulse (710), the centroid glottal pulses and the distance metric;
d. determining Eigenvectors of the vector database (715); and
e. forming parametric models by associating a glottal pulse (710) from the glottal pulse database (840) to each determined Eigenvector.

2. The method of claim 1, further comprising the step of training the formed parametric models for use in speech synthesis;
wherein the training further comprises the steps of:

a. defining a training text corpus;
b. obtaining speech data by recording a speaker speaking the training text;
c. converting the training text into context dependent phone labels;
d. determining the spectral features of the speech data using the phone labels;
e. estimating the fundamental frequency of the speech data; and
f. performing parameter estimation on an audio stream from the obtained speech data using the spectral features and the fundamental frequency.

3. The method of claim 1, wherein the number of glottal pulses (710) is two.

4. The method of claim 1, wherein step (a) further comprises the steps of:

a. de-composing the number of glottal pulses (710) into corresponding sub-band components;
b. computing a sub-band distance metric between the corresponding sub-band components of each glottal pulse (710); and
c. computing the glottal pulse distance metric mathematically using the sub-band distance metrics.

5. The method of claim 4, wherein the computing of step (c) is performed using the mathematical equation:

$$d(x_i, y_i) = \sqrt{d_s^2\left(x_i^{(1)}, y_i^{(1)}\right) + d_s^2\left(x_i^{(2)}, y_i^{(2)}\right) + d_s^2\left(x_i^{(3)}, y_i^{(3)}\right)}$$

Where $d(x_i,y_i)$ represents the distance metric and $d_s^2\left(x_i^{(n)}, y_i^{(n)}\right)$ represents the sub-band distance metrics.

6. The method of claim 1, wherein the number of clusters is 256.

7. The method of claim 1, wherein the clustering of step (b) is performed using a modified k-means calculation that utilizes the glottal pulse distance metric.

8. The method of claim 7, wherein the modified k-means calculation further comprises updating a centroid of a cluster with an element of the cluster whose sum of squares of distances from all other elements of that cluster is minimum.

9. The method of claim 8, further comprising terminating the clustering iterations when there is no shift in any of the centroids from the clusters.

10. The method of claim 1, wherein the determining of Eigenvectors of step (d) is performed using Principal Component Analysis.

11. The method of claim 1, wherein step (e) further comprises the steps of:

a. determining an Eigenvector;
b. determining the closest matching vector (730) from the vector database (715) to the Eigenvector;
c. determining the closest matching glottal pulse from the glottal pulse database; and
d. naming the glottal pulse (710) from the glottal pulse database (840) that is the closest match to the Eigenvector as the Eigen glottal pulse (735) associated with the Eigenvector.

Patentansprüche

1. Verfahren (700) zum Erzeugen parametrischer Modelle, die folgenden Schritte umfassend:

a.Berechnen einer Glottalimpuls-Abstandsmetrik zwischen einer Anzahl von Glottalimpulsen (710);
b. Clustern einer Glottalimpuls-Datenbank (840), die die Glottalimpulse (710) umfasst, in eine Anzahl von Clustern, um Schwerpunkt-Glottalimpulse zu bestimmen;
c. Erzeugen einer entsprechenden Vektordatenbank (715) durch Zuordnen eines Vektors zu jedem Glottalimpuls (710) in der Glottalimpuls-Datenbank (840), wobei der jedem Glottalimpuls (710) zugeordnete Vektor basierend auf dem Glottalimpuls (710), den Schwerpunkt-Glottalimpulsen und der Abstandsmetrik definiert ist;
d. Bestimmen von Eigenvektoren der Vektordatenbank (715) ; und
e. Erzeugen parametrischer Modelle durch Zuordnen eines Glottalimpulses (710) aus der Glottalimpuls-Datenbank (840) zu jedem bestimmten Eigenvektor.

2. Verfahren nach Anspruch 1, ferner umfassend den Schritt des Trainierens der erzeugten parametrischen Modelle zur Verwendung in der Sprachsynthese;
wobei das Trainieren ferner die folgenden Schritte umfasst:

a.Definieren eines Trainingstextkorpus;
b. Erhalten von Sprachdaten durch Aufzeichnen eines Sprechers, der den Trainingstext spricht;
c. Konvertieren des Trainingstextes in kontextabhängige Phon-Kennzeichen;
d. Bestimmen der spektralen Merkmale der Sprachdaten unter Verwendung der Phon-Kennzeichen;
e. Schätzen der Grundfrequenz der Sprachdaten; und
f. Durchführen einer Parameterschätzung für einen Audiodatenstrom aus den erhaltenen Sprachdaten unter Verwendung der Spektralmerkmale und der Grundfrequenz.

3. Verfahren nach Anspruch 1, wobei die Anzahl der Glottalimpulse (710) zwei beträgt.

4. Verfahren nach Anspruch 1, wobei Schritt (a) ferner die folgenden Schritte umfasst:

a. Zerlegen der Anzahl von Glottalimpulsen (710) in entsprechende Teilbandkomponenten;

b.Berechnen einer Teilband-Abstandsmetrik zwischen den entsprechenden Teilband-Komponenten jedes Glottalimpulses (710); und
c. mathematisches Berechnen der Glottalimpuls-Abstandsmetrik unter Verwendung der Teilband-Abstandsmetriken.

5. Verfahren nach Anspruch 4, wobei das Berechnen von Schritt (c) unter Verwendung der folgenden mathematischen Gleichung durchgeführt wird:

$$d(x_i, y_i) = \sqrt{d_s^2\left(x_i^{(1)}, y_i^{(1)}\right) + d_s^2\left(x_i^{(2)}, y_i^{(2)}\right) + d_s^2\left(x_i^{(3)}, y_i^{(3)}\right)}$$

Wobei $d(x_i,y_i)$ die Abstandsmetrik darstellt und $d_s^2\left(x_i^{(n)}, y_i^{(n)}\right)$ die Teilband-Abstandsmetriken darstellt.

6. Verfahren nach Anspruch 1, wobei die Anzahl von Clustern 256 beträgt.

7. Verfahren nach Anspruch 1, wobei das Clustern von Schritt (b) unter Verwendung einer modifizierten k-Means-Berechnung durchgeführt wird, die die Glottalimpuls-Abstandsmetrik nutzt.

8. Verfahren nach Anspruch 7, wobei die modifizierte k-Means-Berechnung ferner das Aktualisieren eines Schwerpunkts eines Clusters mit einem Element des Clusters umfasst, dessen Summe der Quadrate von Abständen von allen anderen Elementen dieses Clusters minimal ist.

9. Verfahren nach Anspruch 8, ferner umfassend das Beenden der Clusteriterationen, wenn keine Verschiebung in einem der Schwerpunkte von den Clustern vorliegt.

10. Verfahren nach Anspruch 1, wobei das Bestimmen der Eigenvektoren von Schritt (d) unter Verwendung der Hauptkomponentenanalyse durchgeführt wird.

11. Verfahren nach Anspruch 1, wobei Schritt (e) ferner die folgenden Schritte umfasst:

a. Bestimmen eines Eigenvektors;
b. Bestimmen des am besten mit dem Eigenvektor übereinstimmenden Vektors (730) aus der Vektordatenbank (715) ;
c. Bestimmen des am besten übereinstimmenden Glottalimpulses aus der Glottalimpuls-Datenbank; und
d. Benennen des Glottalimpulses (710) aus der Glottalimpuls-Datenbank (840), der die beste Übereinstimmung mit dem Eigenvektor aufweist, als der dem Eigenvektor zugeordnete Eigen-Glottalimpuls (735).

Revendications

1. Procédé (700) pour former des modèles paramétriques, comprenant les étapes suivantes :

a. le calcul d'une métrique de distance d'impulsion glottale entre un certain nombre d'impulsions glottales (710) ;
b. le regroupement d'une base de données d'impulsions glottales (840) comprenant les impulsions glottales (710) en un certain nombre de groupes pour déterminer des impulsions glottales centroïdes ;
c. la formation d'une base de données de vecteurs correspondante (715) par l'association d'un vecteur avec chaque impulsion glottale (710) dans la base de données d'impulsions glottales (840), dans lequel le vecteur associé à chaque impulsion glottale (710) est défini sur la base de l'impulsion glottale (710), des impulsions glottales centroïdes et de la métrique de distance ;
d. la détermination de vecteurs propres de la base de données de vecteurs (715) ; et
e. la formation de modèles paramétriques par l'association d'une impulsion glottale (710) issue de la base de données d'impulsions glottales (840) avec chaque vecteur propre déterminé.

2. Procédé selon la revendication 1, comprenant en outre l'étape d'entraînement des modèles paramétriques formés pour une utilisation en synthèse vocale ;
dans lequel l'entraînement comprend en outre les étapes suivantes :

a. la définition d'un corpus de textes d'entraînement ;
b. l'obtention de données vocales par l'enregistrement d'un orateur prononçant le texte d'entraînement ;
c. la conversion du texte d'entraînement en marqueurs phoniques dépendant du contexte ;
d. la détermination des caractéristiques spectrales des données vocales à l'aide des marqueurs phoniques ;
e. l'estimation de la fréquence fondamentale des données vocales ; et
f. la réalisation d'une estimation de paramètre sur un flux audio issu des données vocales obtenues à l'aide des caractéristiques spectrales et de la fréquence fondamentale.

3. Procédé selon la revendication 1, dans lequel le nombre d'impulsions glottales (710) est deux.

4. Procédé selon la revendication 1, dans lequel l'étape (a) comprend en outre les étapes suivantes :

a. la décomposition du nombre d'impulsions glottales (710) en composantes de sous-bande correspondantes ;
b. le calcul d'une métrique de distance de sous-bande entre les composantes de sous-bande correspondantes de chaque impulsion glottale (710) ; et
c. le calcul de la métrique de distance d'impulsion glottale mathématiquement à l'aide des métriques de distance de sous-bande.

5. Procédé selon la revendication 4, dans lequel le calcul de l'étape (c) est réalisé à l'aide de l'équation mathématique :

$$d(x_i, y_i) = \sqrt{d_s^2\left(x_i^{(1)}, y_i^{(1)}\right) + d_s^2\left(x_i^{(2)}, y_i^{(2)}\right) + d_s^2\left(x_i^{(3)}, y_i^{(3)}\right)}$$

où $d(x_i,y_i)$ représente la métrique de distance et $d_s^2\left(x_i^{(n)}, y_i^{(n)}\right)$ représente les métriques de distance de sous-bande.

6. Procédé selon la revendication 1, dans lequel le nombre de groupes est 256.

7. Procédé selon la revendication 1, dans lequel le regroupement de l'étape (b) est réalisé à l'aide d'un calcul à k-moyennes modifié qui utilise la métrique de distance d'impulsion glottale.

8. Procédé selon la revendication 7, dans lequel le calcul à k-moyennes modifié comprend en outre la mise à jour d'un centroïde d'un groupe avec un élément du groupe dont la somme des carrés des distances par rapport à tous les autres éléments de ce groupe est minimale.

9. Procédé selon la revendication 8, comprenant en outre la terminaison des itérations de regroupement lorsqu'il n'y a aucun décalage dans l'un quelconque des centroïdes issus des groupes.

10. Procédé selon la revendication 1, dans lequel la détermination de vecteurs propres de l'étape (d) est réalisée à l'aide d'une analyse en composantes principales.

11. Procédé selon la revendication 1, dans lequel l'étape (e) comprend en outre les étapes suivantes :

a. la détermination d'un vecteur propre ;
b. la détermination du vecteur ayant la concordance la plus proche (730) issu de la base de données de vecteurs (715) avec le vecteur propre ;
c. la détermination de l'impulsion glottale ayant la concordance la plus proche issue de la base de données d'impulsions glottales ; et
d. la dénomination de l'impulsion glottale (710) issue de la base de données d'impulsions glottales (840) qui est concordance la plus proche du vecteur propre comme l'impulsion glottale propre (735) associée au vecteur propre.

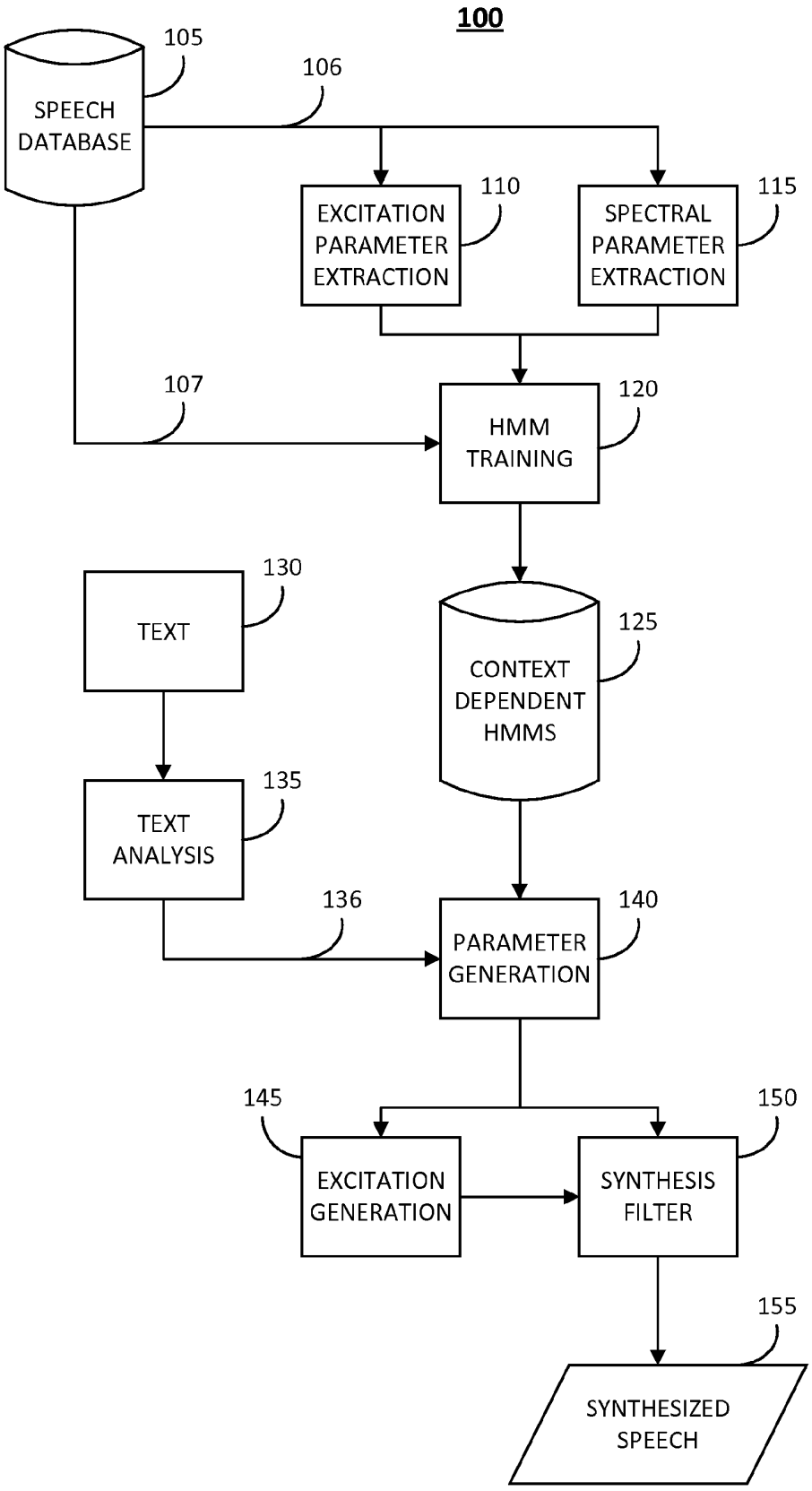
100
105
SPEECH DATABASE
106
EXCITATION PARAMETER EXTRACTION
110
SPECTRAL PARAMETER EXTRACTION
115
107
HMM TRAINING
120
130
TEXT
CONTEXT DEPENDENT HMMS
125
TEXT ANALYSIS
135
136
PARAMETER GENERATION
140
145
EXCITATION GENERATION
SYNTHESIS FILTER
150
155
SYNTHESIZED SPEECH

FIG. 1

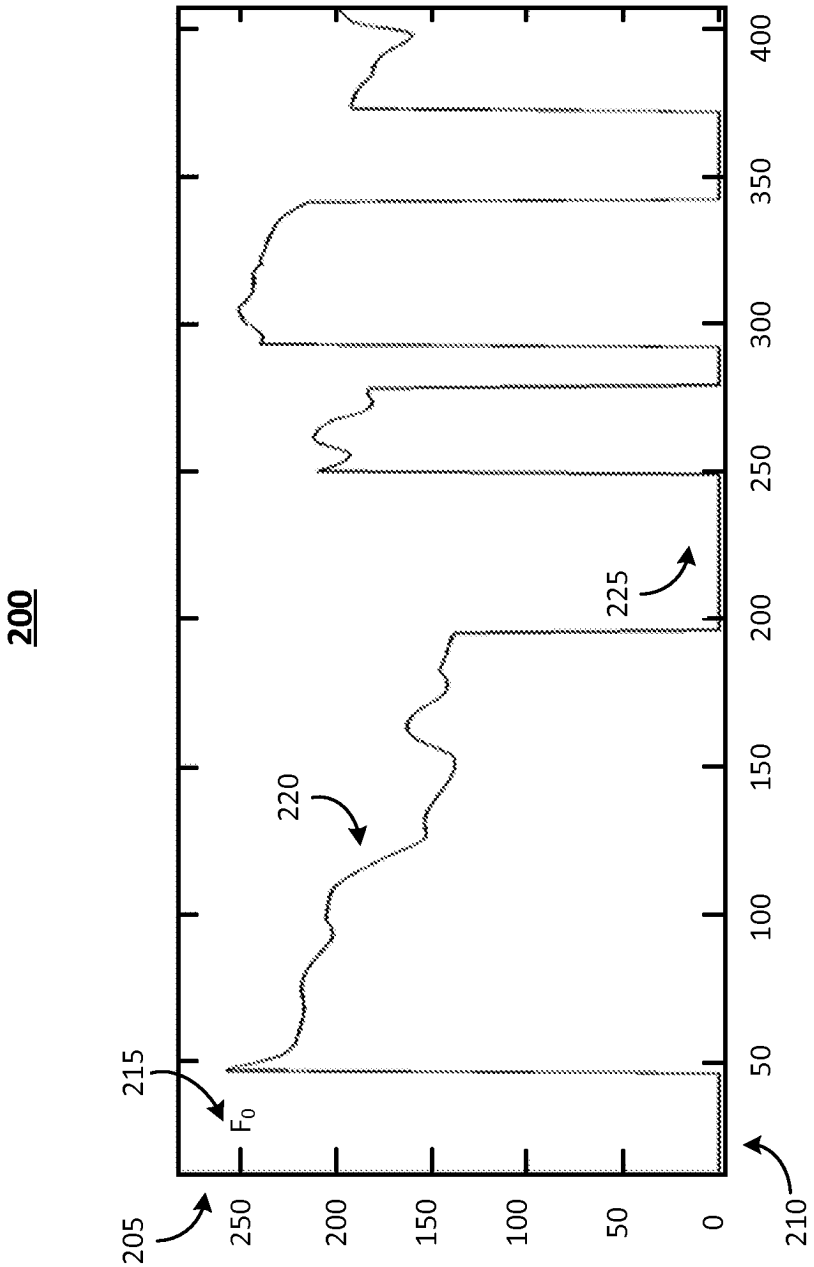
200
205
215
F_0
220
225
210
250
200
150
100
50
0
50
100
150
200
250
300
350
400

FIG. 2

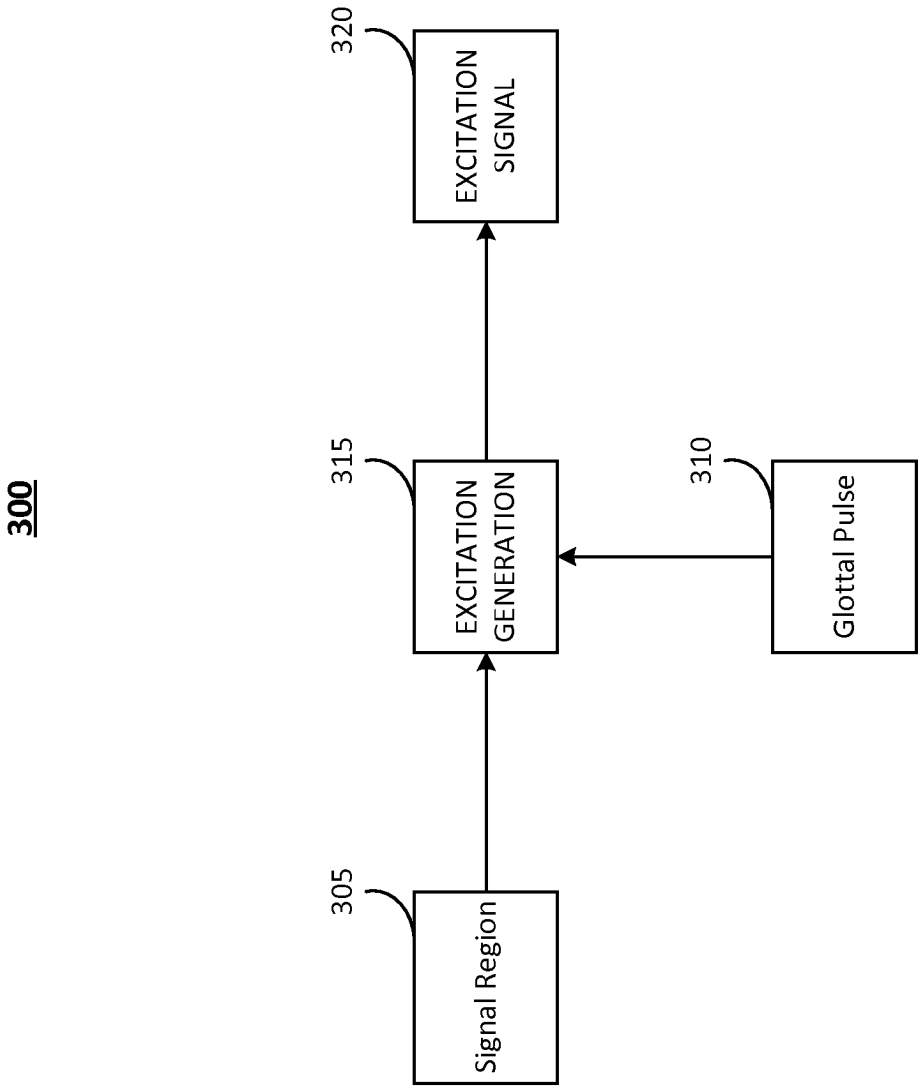
300
320
EXCITATION SIGNAL
315
EXCITATION GENERATION
310
Glottal Pulse
305
Signal Region

FIG. 3

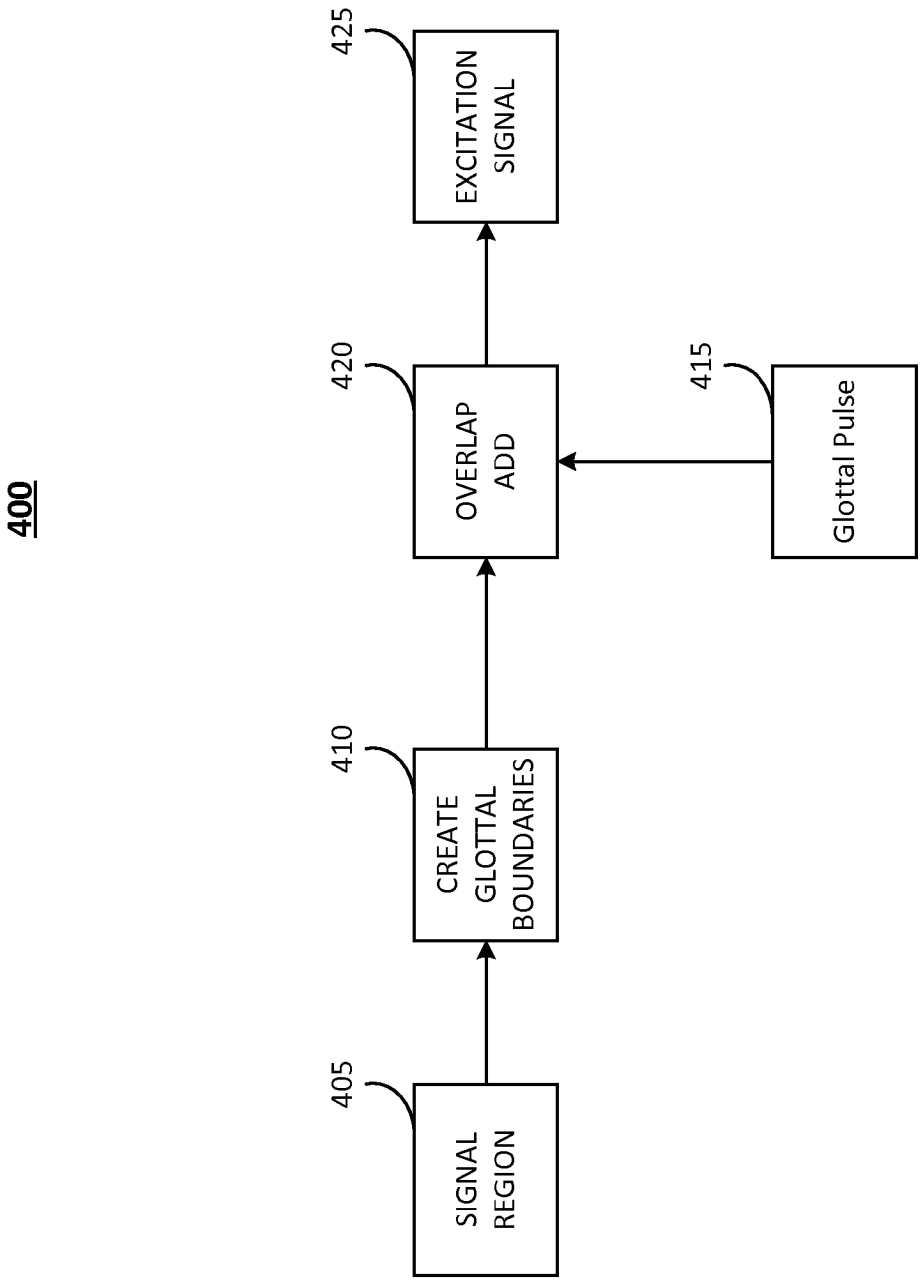
400
405
SIGNAL REGION
410
CREATE GLOTTAL BOUNDARIES
420
OVERLAP ADD
415
Glottal Pulse
425
EXCITATION SIGNAL

FIG. 4

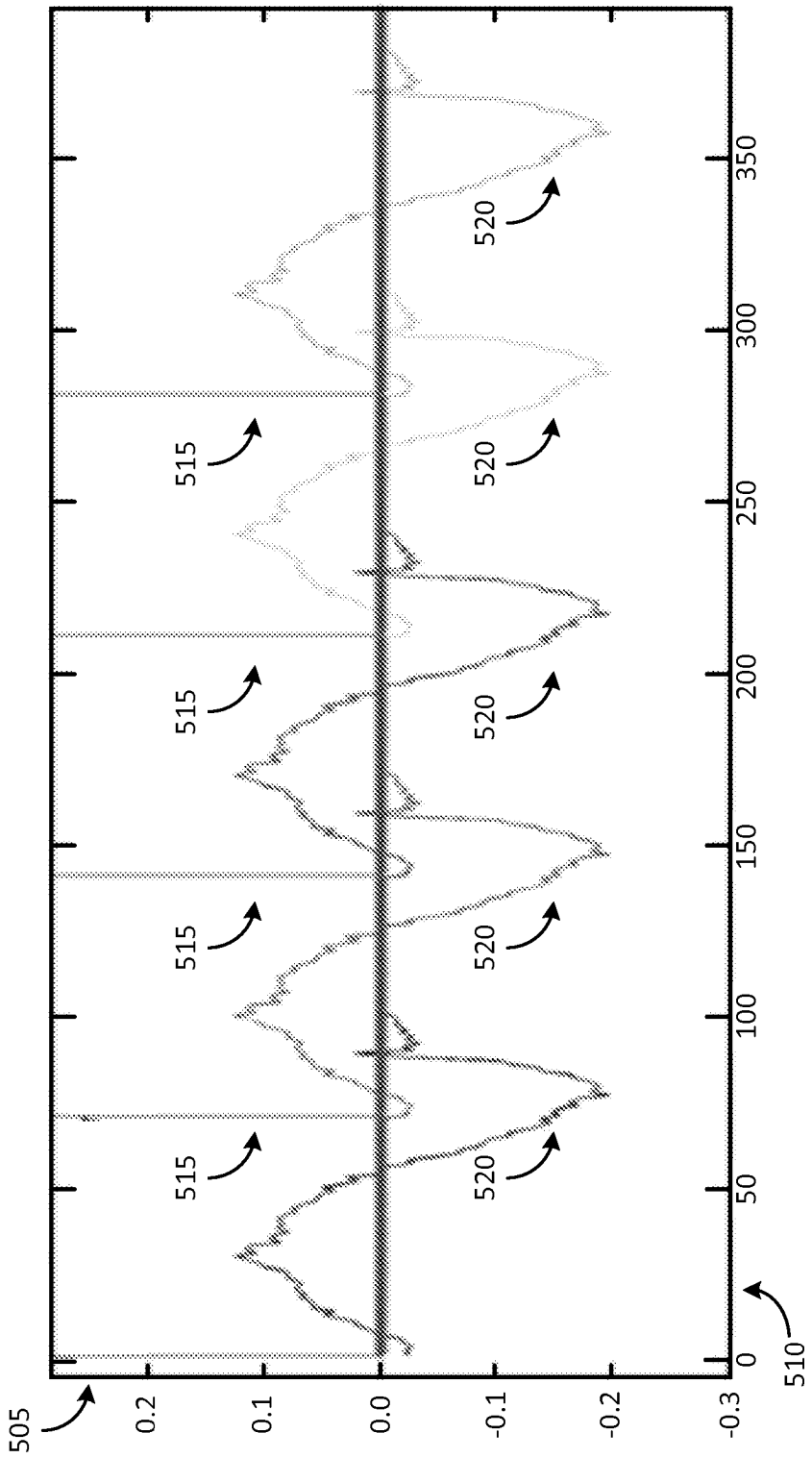

FIG. 5

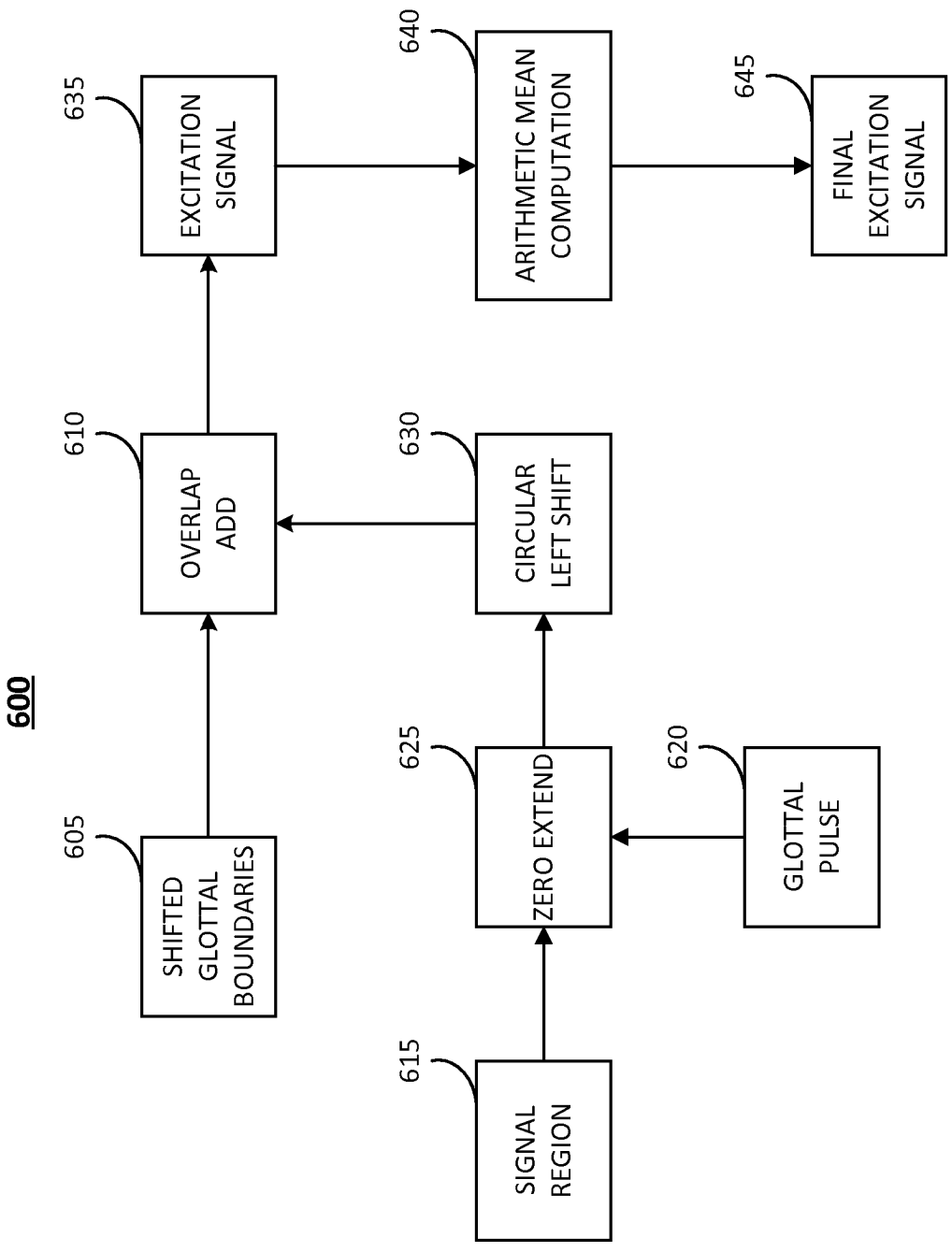
600
605
SHIFTED GLOTTAL BOUNDARIES
610
OVERLAP ADD
615
SIGNAL REGION
620
GLOTTAL PULSE
625
ZERO EXTEND
630
CIRCULAR LEFT SHIFT
635
EXCITATION SIGNAL
640
ARITHMETIC MEAN COMPUTATION
645
FINAL EXCITATION SIGNAL

FIG. 6

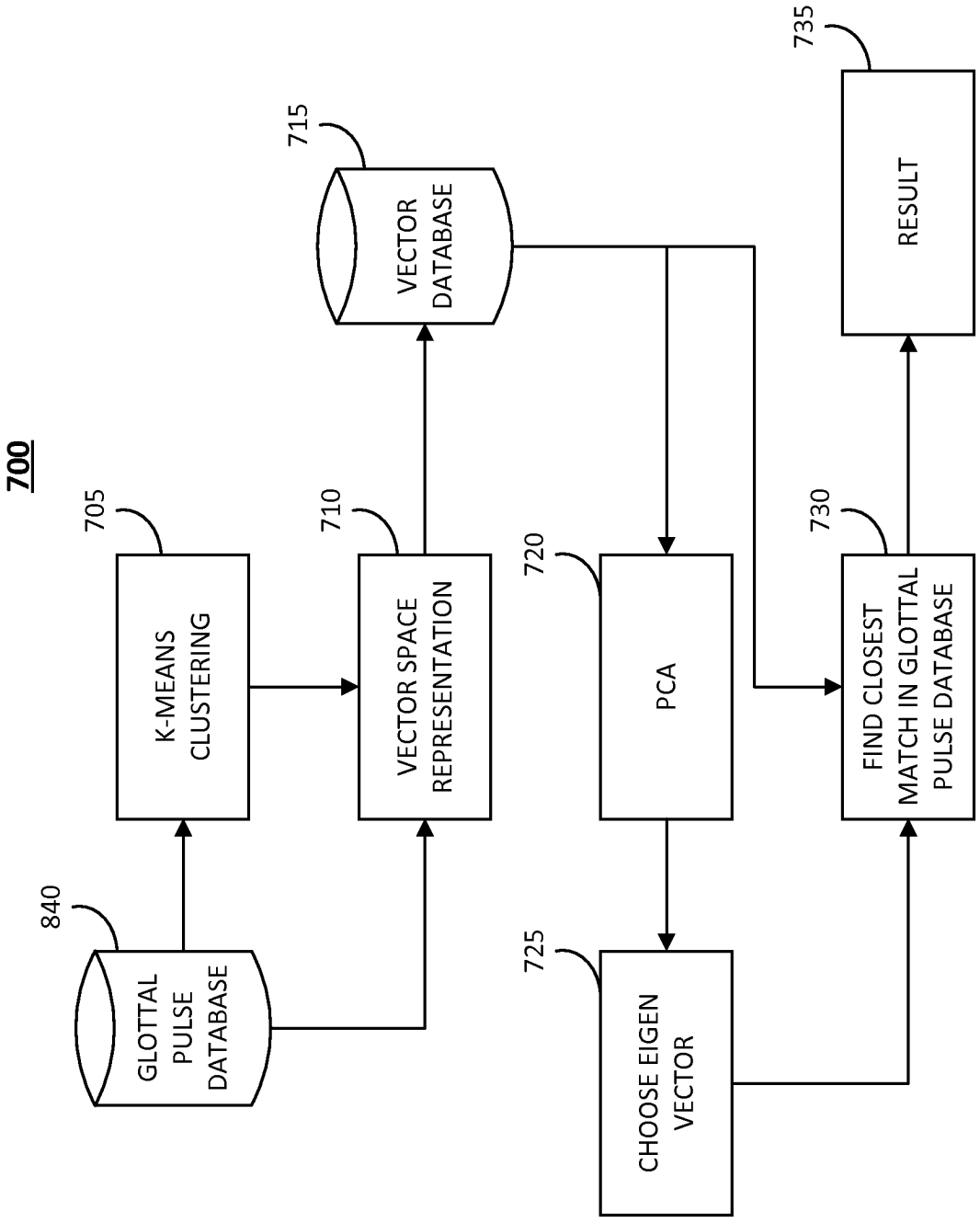
700
705
K-MEANS CLUSTERING
710
VECTOR SPACE REPRESENTATION
715
VECTOR DATABASE
720
PCA
725
CHOOSE EIGEN VECTOR
730
FIND CLOSEST MATCH IN GLOTTAL PULSE DATABASE
735
RESULT
840
GLOTTAL PULSE DATABASE

FIG. 7

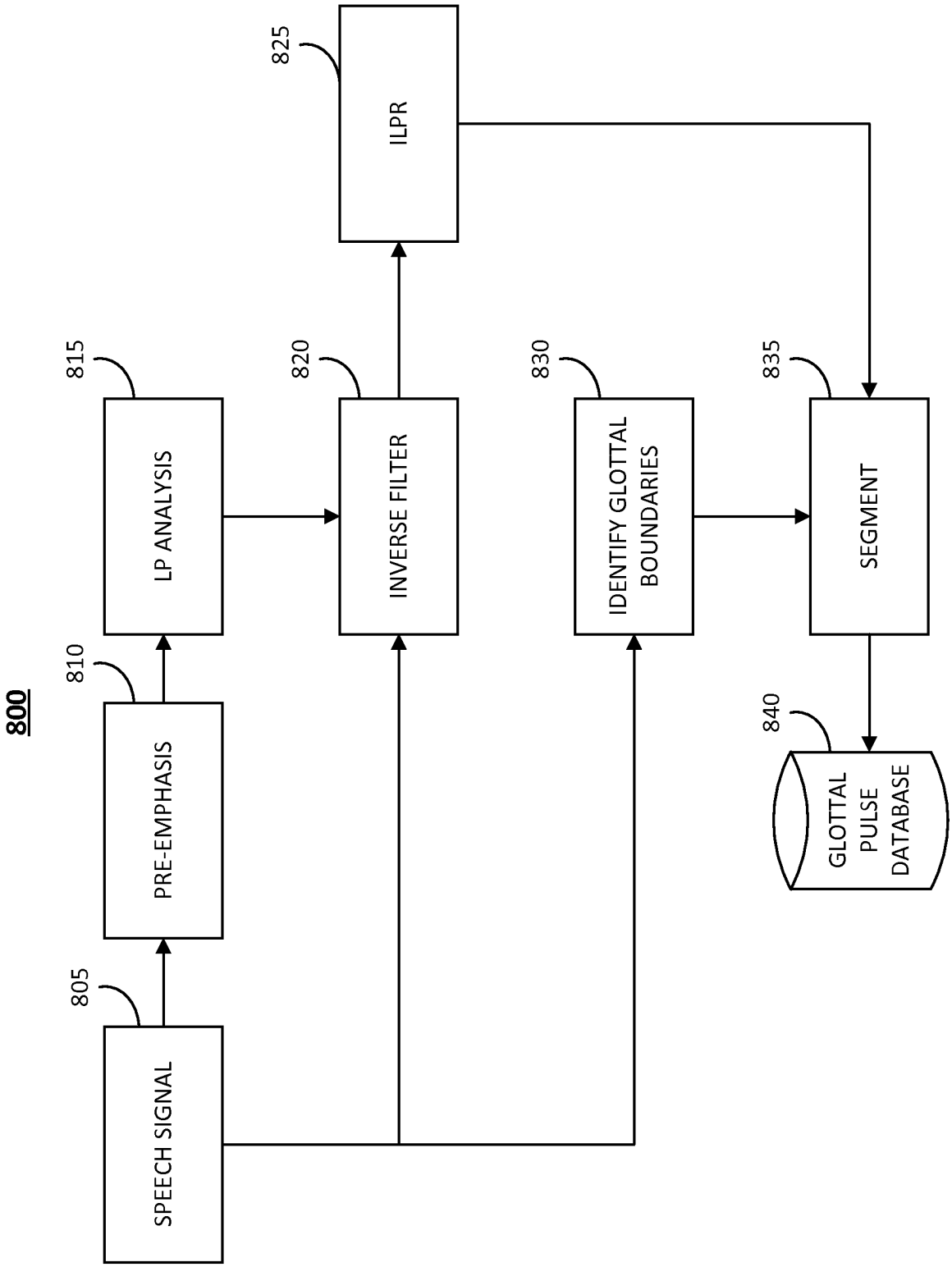
800
805
SPEECH SIGNAL
810
PRE-EMPHASIS
815
LP ANALYSIS
820
INVERSE FILTER
825
ILPR
830
IDENTIFY GLOTTAL BOUNDARIES
835
SEGMENT
840
GLOTTAL PULSE DATABASE

FIG. 8

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

- EP 2242045 A, Thomas Drugman **[0007]**

Non-patent literature cited in the description

- Comparing Glottal-Flow-Excited Statistical Parametric Speech Synthesis Methods. 2103 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Institute of Electrical and Electronics Engineers, 26 May 2013, 7830-7834 **[0002]**
- A Novel Codebook-Based Excitation Model for us in Speech Synthesis. Cognitive Info Communications (COGINFOCOM), 2012 IEEE 3rd International Conference on. IEEE, 02 December 2012, 661-665 **[0002]**