



(51) International Patent Classification:
G10L 13/02 (2006.01)

(21) International Application Number:
PCT/EP2010/059416

(22) International Filing Date:
1 July 2010 (01.07.2010)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
0911494.3 2 July 2009 (02.07.2009) GB

(71) Applicant (for all designated States except US): **THE UNIVERSITY COURT OF THE UNVIERSITY OF EDINBURGH** [GB/GB]; Old College, Edinburgh EH8 9YL (GB).

(72) Inventors; and

(75) Inventors/Applicants (for US only): **YAMAGISHI, Junichi** [JP/GB]; Centre for Speech Technology Research, The University of Edinburgh, School of Informatics, Informatics Forum, 10 Crichton Street Edinburgh EH8 9AB (GB). **KING, Simon, Alistair** [GB/GB]; Centre for Speech Technology Research, The University of Edinburgh, School of Informatics, Informatics Forum, 10 Crichton Street Edinburgh EH8 9AB (GB). **RENALS, Stephen, John** [GB/GB]; Centre for Speech Technology Research, The University of Edinburgh, School of Informatics, Informatics Forum, 10 Crichton Street Edinburgh EH8 9AB (GB).

(74) Agents: **HIGGIN, Paul** et al.; 48 Friar Gate, Derby, Derbyshire DE1 1GY (GB).

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

(84) Designated States (unless otherwise indicated, for every kind of regional protection available): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Published:

- with international search report (Art. 21(3))
- before the expiration of the time limit for amending the claims and to be republished in the event of receipt of amendments (Rule 48.2(h))

(54) Title: ENABLING SYNTHESIS OF SPEECH HAVING A TARGET CHARACTERISTIC

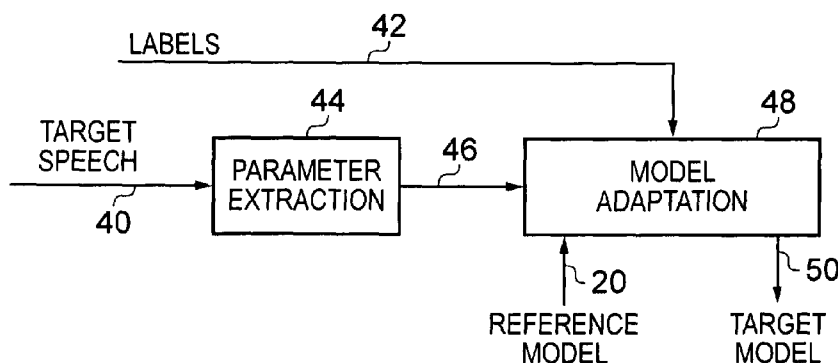


FIG. 3

(57) Abstract: A method involving: inputting target speech features derived from speech having a target characteristic; inputting labelling of the speech; and adapting, using a speech feature transform, statistical parameters of a predefined statistical parametric speech model to create an adapted statistical parametric speech model configured to synthesis speech having the target characteristic.

TITLE

Enabling synthesis of speech having a target characteristic.

5 TECHNOLOGICAL FIELD

Embodiments of the present invention relate to enabling synthesis of speech having a target characteristic.

10 BACKGROUND

It is desirable to be able to use a machine to produce speech from non-speech input without having to record the actual speech produced. The most common application of speech synthesis is in 'text to speech' in which
15 the input is written language and the output is a speech signal.

Most commercial systems for synthesising speech use a method called 'unit selection' that relies on a large database of several hours of speech from a single human speaker (the 'target speaker'), typically recorded by a
20 professional in a studio on high quality equipment under carefully controlled conditions, meticulously labelled and carefully checked. To synthesise new sentences, small fragments of speech from the database are selected using an automatic procedure, concatenated together and played back. This works very well, but is expensive and inflexible. In particular, creating a new voice for
25 such a system requires considerable expertise and labour, and therefore costs a substantial amount of money and is well beyond the reach of individual users.

BRIEF SUMMARY

30

An alternative to unit selection is a method which also uses a database of recorded speech, but this time to construct a statistical model of speech from which any required new sentence can be generated. The statistical model comprises a number of statistical parameters which collectively describe the

characteristics of speech. The model allows the generation of new words and phrases not contained in the original database. Because the model is statistical, it is more tolerant of certain problems or shortcomings in the data. It does not require such careful labelling or checking of the data and can give
5 good performance with a smaller database of about one hour of speech from the target speaker.

According to some but not necessarily all embodiments of the invention there is provided a method comprising: inputting target speech features derived
10 from speech having a target characteristic; inputting labelling of the speech; and adapting, using a speech feature transform, statistical parameters of a predefined statistical parametric speech model to create an adapted statistical parametric speech model configured to synthesis speech having the target characteristic.

15

According to some but not necessarily all embodiments of the invention there is provided an apparatus comprising: means for producing target speech features from speech having a target characteristic; means for receiving labelling of the speech; and means for adapting, using a speech feature
20 transform, statistical parameters of a predefined statistical parametric speech model to create an adapted statistical parametric speech model configured to synthesis speech having the target characteristic.

According to some but not necessarily all embodiments of the invention there is provided an apparatus comprising: adaption circuitry configured to adapt
25 statistical parameters of a predefined statistical parametric speech model to create an adapted statistical parametric speech model configured to synthesise speech having a target characteristic, wherein the adaptation of the statistical parameters uses a transform to transform speech features
30 derived from speech having a target characteristic.

Embodiments of the invention may allow the creation of new voices using just a few minutes of recorded speech from the target speaker. Embodiments may also utilise noisy or otherwise imperfect recordings. The parameters of the

statistical parametric speech model are automatically adapted, using a transform of speech features, to match the characteristics of speech from the target speaker.

5 BRIEF DESCRIPTION

For a better understanding of various examples of embodiments of the present invention reference will now be made by way of example only to the accompanying drawings in which:

10

Fig 1 schematically illustrates an example of a system and method for predefining statistical parametric models of speech;

Fig 2 schematically illustrates an example of a system and method for producing synthesised speech using the predefined statistical parametric

15 models of speech;

Fig 3 schematically illustrates an example of a system and method that adapts existing statistical parametric models of speech to create an adapted statistical parametric model of speech that is optimized to create target synthetic speech;

20 Fig 4 schematically illustrates an example of a system and method for producing synthesised speech using the adapted statistical parametric models of speech;

Fig 5 schematically illustrates one example of a method that may occur at the model training block of Fig 1;

25 Fig 6 schematically illustrates the operation of the model adaptation block of Fig 3;

Fig 7 schematically illustrates one example of the method 48 that may occur at the model adaptation block Fig 3; and

30 Fig 8 schematically illustrates an apparatus suitable for providing the aforementioned systems and methods.

DETAILED DESCRIPTION

Fig 1 schematically illustrates a system and method for predefining statistical parametric models of speech (reference models) 20.

Annotated reference speech includes the reference speech 10 and labels 12 for the reference speech that provide a linguistic specification of the speech. The reference speech 10 is converted, by parameter extraction block 14, into parameters in the feature space domain. These parameters which may also be called speech features 16 are provided as one input to a model training block 18. Another input to the model training block 18 is the labels 12.

The model training block 18, as is known in the art, determines statistical parametric models of speech (reference models) 20 that have the maximum likelihood of producing the reference speech 10 from the labels 12. Typically the model training block 18 uses machine learning or training to determine the statistical parametric models of speech (reference models) 20.

Fig 2 schematically illustrates a system and method for producing synthesised speech 34 using the predefined statistical parametric models of speech (reference models) 20.

Text 22 is analysed at block 24 to produce labels 26 that provide a linguistic specification. The labels 26 and the reference model 20 are inputs to speech generator block 28. The speech generator block 28 may use labels 26 to select appropriate reference models 20 for speech generation. The selected reference models are then used to generate parameters 30 from the labels 26. These parameters may also be called speech features 30.

The speech features are input into a speech synthesis engine 32 which generates synthetic speech 34.

Typically the block 14 that performs parameter extraction converts the reference speech 10 from the time domain to the frequency domain. It may for example be a suitable voice coder. Some voice coders produce as outputs excitation parameters and spectral parameters, for example.

The reference speech 10 is typically carefully produced in a studio and is comprehensive in its inclusion of phonemes so that the reference models 20 are accurate and comprehensive. Creating and annotating the reference
 5 speech 10 may be expensive and time consuming.

One or more statistical parametric models of speech (reference models) 20 may be defined for each different phoneme. One suitable statistical parametric model is a Hidden State Statistical Model, such as a Hidden Markov Model
 10 (HMM) or Hidden Semi-Markov Model (HSMM), which is defined using statistical parameters. Each hidden state has at least one probability distribution for the possible output values of the state. The model may therefore associate each possible output vector with a probability distribution, defined using statistical parameters. If the distribution is Gaussian, the
 15 distribution can be defined by the using the mean and variance as the statistical parameters. A HMM is defined by initial state probabilities, state transition probabilities, and state output probability distributions. A Hidden Semi Markov Model (HSMM) may have state duration probability distribution instead of or in addition to state transition probabilities.

20 If an N -state left-to-right HSMM λ with no skip paths is specified by a state i output probability distribution b_i and a state duration probability distribution p_i then the probability distribution b_i is dependent upon the output observation vector and statistical parameters (mean vector and covariance matrix) and
 25 the state duration d is dependent upon state duration probability p_i and the statistical parameters (*mean and variance*).

The model training block 18 may determine maximum likelihood estimates of the parameters of the Model given a target output observation vector. The
 30 model is found that maximises the observation probability of the reference speech 10 for the labels 12. This may be achieved using the Baum-Welch algorithm.

Fig 5 schematically illustrates one example of a method that may occur at the model training block 18 of Fig 1. Initially at block 60, the speech features 16, labels 12 and current reference model are time aligned. Then at block 62 the expected values of probability distribution b_i and the state duration probability p_i are determined using the existing statistical parameters.

$$b_i(\mathbf{o}) = N(\mathbf{o}; \boldsymbol{\mu}_i; \boldsymbol{\Sigma}_i)$$

$$p_i(d) = N(d; m_i; \sigma_i^2)$$

10

where

$\mathbf{o}$ is the output observation vector

$\boldsymbol{\mu}_i$ is a mean vector (statistical parameter)

$\boldsymbol{\Sigma}_i$ is a covariance matrix (statistical parameter)

15

d is the state duration

m_i is the mean duration (statistical parameter)

σ_i^2 is the variance of the duration (statistical parameter)

Then at block 64, the statistical parameters are updated. For example, the expected values may be used to re-estimate the statistical parameters of the model. This may, for example, use re-estimation formula based on the Baum-Welch algorithm.

Then at block 66, the model is updated using the expected values determined at block 62 and the re-estimated statistical parameters determined at block 64.

$$\tilde{\lambda} = \underset{\lambda}{\operatorname{argmax}} P(\mathbf{O}|\lambda).$$

where

30

$$P(\mathbf{O}|\lambda) = \sum_{i=1}^N \sum_{\substack{j=1 \\ j \neq i}}^N \sum_{d=1}^t \alpha_{t-d}(j) p_i(d) \prod_{s=t-d+1}^t b_i(\mathbf{o}_s) \beta_t(i)$$

The process is then repeated until there is convergence. The current model
 5 may then be output as the reference model 20.

The inventors have realised that it would be beneficial to somehow use the
 existing statistical parametric models of speech (reference models) 20 for
 creating target synthetic speech with different target characteristics. These
 10 characteristics may include speaker identity, emotion, urgency, gravity, accent
 etc.

Fig 3 schematically illustrates an example of a system and method by which
 this may be achieved. Annotated target speech includes the target speech 40
 15 and labels 42 for the target speech that provide a linguistic specification of the
 target speech. The target speech 40 is converted, by the parameter extraction
 block 44, into parameters in the feature space domain. These parameters
 which may also be called speech features 46 are provided as one input to a
 model adaptation block 48. Another input to the model adaptation block 48 is
 20 the labels 42.

The model adaptation block 48 determines target statistical parametric models
 of speech (target models) 50 that have the maximum likelihood of producing
 the target speech 40 from the labels 42. The target statistical parametric
 25 models of speech (target models) 50 are based upon the reference models
 20.

Fig 4 schematically illustrates a method for producing synthesised speech 54
 using the adapted statistical parametric models of speech (target models) 50.

30

Text 22 is analysed at block 24 to produce labels 26 that provide a linguistic specification. The labels 26 and the reference model 20 are input to speech generator block 28. The speech generator block 28 may use a label 26 to select an appropriate one of the target models 50 for speech generation. The selected model is then used to generate parameters 52. These parameters may also be called speech features 30.

The speech features 52 are input into a speech synthesis engine 32 which generates synthetic target speech 54.

10

Fig 6 schematically illustrates the operation of the model adaptation block 48. The model adaptation block 48 determines speech feature transforms 76 that transform the speech features 74 defined in the reference models. Each phoneme 70_j has its own associated reference model 20_j . A transform 76_j transforms the speech feature 74_j of the reference model 20_j to the target speech features 72_j . These speech feature transforms 76_j are used to create target models 50_j as illustrated in Fig 7, The transforms 76_j are determined, in speech feature space, that adapt the reference model 20 such that the maximum likelihood is that the adapted model (target model 50) creates as output the target speech features 46.

20

The transforms 76 minimise the differences between the target output 72 and the model output 74 by transforming the output observation vector of the reference model 20 in feature space as part of the maximum likelihood determination that estimates the maximum likelihood statistical parameters of the target model 50 given the target output vector (target speech features 46).

25

As an example, if an N -state left-to-right HSMM λ with no skip paths is specified by a state i output probability distribution b_i and a state duration probability distribution p_i then the probability distribution b_i is dependent upon the output observation vector o and statistical parameters (mean vector and covariance matrix) and the state duration d is dependent upon state duration probability p_i and the statistical parameters (*mean and variance*).

30

Then a first transform may be applied to the output observation vector $\mathbf{o}$ and a second transform may be applied to the state duration $\mathbf{d}$.

For example, the observation vector may undergo a linear transform such as:

5

$$\mathbf{o} \rightarrow \boldsymbol{\xi} \mathbf{o} + \boldsymbol{\varepsilon}$$

where

$\boldsymbol{\xi}$ is a matrix that can scale, rotate and shear

$\boldsymbol{\varepsilon}$ is a vector that shifts

10

This results in the expected values of the probability distribution b_i being expressed as

$$b_i(\mathbf{o}) = |\boldsymbol{\xi}| \mathcal{N}(\boldsymbol{\xi} \mathbf{o} + \boldsymbol{\varepsilon}; \boldsymbol{\mu}_i; \boldsymbol{\Sigma}_i)$$

15

For example, the state duration may undergo a linear transform such as:

$$\mathbf{d} \rightarrow \chi \mathbf{d} + \mathbf{v}$$

20 where

χ is a scalar

$\mathbf{v}$ is a shift

This results in the expected values of the probability distribution p_i being

25 expressed as

$$p_i(\mathbf{d}) = |\chi| \mathcal{N}(\chi \mathbf{d} + \mathbf{v}; \mathbf{m}_i; \sigma_i^2)$$

30 Fig 7 schematically illustrates one example of the method 48 that may occur at the model adaptation block 48 of Fig 3.

Initially at block 82, the original speech features 46, labels 42 and the reference model 20 are time aligned. The reference model 20 is predetermined as previously described with reference to Fig 1 and/or Fig5.

- 5 Next at block 82 the expected values of probability distribution b_i and the state duration probability p_i are determined using the existing statistical parameters (μ_i ; Σ_i and m_i ; σ_i^2).

$$b_i(\mathbf{o}) = \frac{1}{|\xi|} \sum_{\xi} N(\xi \mathbf{o} + \epsilon; \mu_i; \Sigma_i)$$

10

$$p_i(d) = \frac{1}{|\chi|} \sum_{\chi} N(\chi d + v; m_i; \sigma_i^2)$$

- Then at block 84, the statistical parameters are updated. For example, the expected values may be used to re-estimate the statistical parameters of the model. This may, for example, use re-estimation formula based on the Baum-Welch algorithm.
- 15

The probability density functions are given by:

20

$$\begin{aligned}
\bar{\mu}_i &= \frac{\sum_{f=1}^F \sum_{t=1}^{T_f} \sum_{d=1}^t \gamma_t^d(i) \sum_{s=t-d+1}^t (\zeta_f \mathbf{o}_{s_f} + \epsilon_f)}{\sum_{f=1}^F \sum_{t=1}^{T_f} \sum_{d=1}^t d \cdot \gamma_t^d(i)} \\
\bar{\Sigma}_i &= \frac{\sum_{f=1}^F \sum_{t=1}^{T_f} \sum_{d=1}^t \gamma_t^d(i) \sum_{s=t-d+1}^t (\zeta_f \mathbf{o}_{s_f} + \epsilon_f - \bar{\mu}_i)(\zeta_f \mathbf{o}_{s_f} + \epsilon_f - \bar{\mu}_i)^\top}{\sum_{f=1}^F \sum_{t=1}^{T_f} \sum_{d=1}^t d \cdot \gamma_t^d(i)} \\
\bar{m}_i &= \frac{\sum_{f=1}^F \sum_{t=1}^{T_f} \sum_{d=1}^t \gamma_t^d(i) \cdot (\chi_f d + \nu_f)}{\sum_{f=1}^F \sum_{t=1}^{T_f} \sum_{d=1}^t \gamma_t^d(i)} \\
\bar{\sigma}_i^2 &= \frac{\sum_{f=1}^F \sum_{t=1}^{T_f} \sum_{d=1}^t \gamma_t^d(i) \cdot (\chi_f d + \nu_f - \bar{m}_i)^2}{\sum_{f=1}^F \sum_{t=1}^{T_f} \sum_{d=1}^t \gamma_t^d(i)}
\end{aligned}$$

where F is the number of training speakers, $\mathbf{O} = \{\mathbf{O}_1, \dots, \mathbf{O}_F\}$ is all the training data, and $\mathbf{O}_f = \{\mathbf{o}_{1f}, \dots, \mathbf{o}_{Tf}\}$ is training data of length T_f for speaker f , where

$$5 \quad \mathbf{o}_{tf} \rightarrow \xi_f \mathbf{o}_{tf} + \epsilon_f \text{ and } \chi_f \rightarrow \chi_f d + \nu_f$$

and where $\gamma_t^d(i)$ is the state occupancy probability of being in state i of the HSMM for the period of time from $t-d+1$ to t given $\mathbf{O}_f$ and is defined as

10

$$\gamma_t^d(i) = \frac{1}{P(\mathbf{O}_f | \lambda)} \sum_{\substack{j=1 \\ j \neq i}}^N \alpha_{t-d}(j) a_{ji} p_i(d) \prod_{s=t-d+1}^t b_i(\mathbf{o}_{s_f}) \beta_t(i).$$

Here, the observation probability of the training data $\mathbf{O}_f$ given the model λ ,

15 $P(\mathbf{O}_f | \lambda)$ and the forward and backward probabilities, $\alpha_t(i)$ $\beta_t(i)$, can be written as

$$\begin{aligned}
P(O_f|\lambda) &= \sum_{i=1}^N \sum_{\substack{j=1 \\ j \neq i}}^N \sum_{d=1}^t \alpha_{t-d}(j) a_{ji} p_i(d) \prod_{s=t-d+1}^t b_i(o_{s_f}) \beta_t(i) \\
\alpha_t(i) &= \sum_{d=1}^t \sum_{\substack{j=1 \\ j \neq i}}^N \alpha_{t-d}(j) a_{ji} p_i(d) \prod_{s=t-d+1}^t b_i(o_{s_f}) \\
\beta_t(i) &= \sum_{d=1}^{T_f-t} \sum_{\substack{j=1 \\ j \neq i}}^N a_{ij} p_j(d) \prod_{s=t+1}^{t+d} b_j(o_{s_f}) \beta_{t+d}(j)
\end{aligned}$$

where $\alpha_0(i) = \pi_i$ and $\beta_T(i) = 1$, and π_i is the initial state probability of being state i at time $t=1$.

5

Then at block 88, the feature space transforms are updated using the calculated expected values and the re-estimated statistical parameters.

$$\tilde{\lambda} = \underset{\lambda}{\operatorname{argmax}} P(O|\lambda).$$

10

The process is then repeated until there is convergence. The target model 50 is defined by the final statistical parameters determined at block 86.

15 Adapting existing reference statistical parametric models 20 to create adapted models 50 for creating target synthetic speech is a cost effective way of synthesising target speech using a small amount of input target speech 40.

20 Constraining the maximum likelihood estimate of the parameters of the statistical model using transforms in the speech feature domain allows the same transform to be applied to both the mean and variance of a Gaussian

density function which reduces processing complexity and increase memory efficiency.

5 Constraining the maximum likelihood estimate of the parameters of the statistical model using transforms in the speech domain automatically removes dependency on non-speech content e.g. background noise etc.

Fig 8 schematically illustrates an apparatus 90 suitable for providing the aforementioned systems and methods.

10 A processor 92 is configured to read from and write to the memory 94. The processor 92 may also comprise an output interface 93 via which data and/or commands are output by the processor 92 and an input interface 91 via which data and/or commands are input to the processor 92.

15 The memory 94 stores a computer program 96 comprising computer program instructions that control the operation of the apparatus 90 when loaded into the processor 92. The computer program instructions 96 provide the logic and routines that enables the apparatus to perform the methods illustrated in the Figs. The processor 92 by reading the memory 94 is able to load and
20 execute the computer program 96.

The computer program may arrive at the apparatus 90 via any suitable delivery mechanism. The delivery mechanism may be, for example, a
25 computer-readable storage medium, a computer program product, a memory device, a record medium such as a CD-ROM or DVD, an article of manufacture that tangibly embodies the computer program. The delivery mechanism may be a signal configured to reliably transfer the computer program.

30 Although the memory 94 is illustrated as a single component it may be implemented as one or more separate components some or all of which may be integrated/removable and/or may provide permanent/semi-permanent/dynamic/cached storage.

The blocks illustrated in the Figs may represent steps in a method and/or sections of code in the computer program and/or functional modules in a system. The illustration of a particular order to the blocks does not necessarily
5 imply that there is a required or preferred order for the blocks and the order and arrangement of the block may be varied. Furthermore, it may be possible for some steps to be omitted.

Although embodiments of the present invention have been described in the
10 preceding paragraphs with reference to various examples, it should be appreciated that modifications to the examples given can be made without departing from the scope of the invention as claimed.

Features described in the preceding description may be used in combinations
15 other than the combinations explicitly described.

Although functions have been described with reference to certain features, those functions may be performable by other features whether described or
20 not.

Although features have been described with reference to certain embodiments, those features may also be present in other embodiments whether described or not.

25 Whilst endeavouring in the foregoing specification to draw attention to those features of the invention believed to be of particular importance it should be understood that the Applicant claims protection in respect of any patentable feature or combination of features hereinbefore referred to and/or shown in the drawings whether or not particular emphasis has been placed thereon.

30

CLAIMS

1. A method comprising:
inputting target speech features derived from speech having a target
5 characteristic;
inputting labelling of the speech; and
adapting, using a speech feature transform, statistical parameters of a
predefined statistical parametric speech model to create an adapted statistical
parametric speech model configured to synthesis speech having the target
10 characteristic.
2. A method as claimed in claim 1 wherein the target speech features are
obtained by parametrically encoding speech recorded by a consumer.
- 15 3. A method as claimed in claim 1 or 2, wherein the target speech features are
obtained by parametrically encoding speech of short duration that is speech of
less than thirty minutes duration and preferably less than approximately ten
minutes duration.
- 20 4. A method as claimed in any preceding claim, wherein the predefined
statistical parametric speech model is a hidden state statistical model.
5. A method as claimed in any preceding claim, wherein the hidden state
statistical model is a Hidden Markov Model or Hidden Semi-Markov Model.
25
6. A method as claimed in any preceding claim, wherein the statistical
parameters define a distribution and where both the mean and variance of any
particular Gaussian are transformed using the same transform.
- 30 7. A method as claimed in any preceding claim, wherein the statistical
parameters include mean and variance or mean vector and co-variance
matrix.

8. A method as claimed in any preceding claim, further comprising determining the transform, in speech feature space, that adapts the predefined statistical parametric speech model such that the adapted statistical parametric speech model creates speech with the target characteristic with maximum likelihood.
9. A method as claimed in any preceding claim, wherein the transforms are linear transforms.
10. A method as claimed in any preceding claim, wherein both the transforms and the adapted statistical parametric speech model are simultaneously trained using the input speech features and the input labelling.
11. A method as claimed in any preceding claim, wherein transforms simultaneously transform both a state output and a state duration.
12. A method comprising synthesising speech using an adapted statistical parametric speech model produced by the method of any one of claims 1 to 11.
13. An apparatus comprising:
means for producing target speech features from speech having a target characteristic;
means for receiving labelling of the speech; and
means for adapting, using a speech feature transform, statistical parameters of a predefined statistical parametric speech model to create an adapted statistical parametric speech model configured to synthesis speech having the target characteristic.
14. An apparatus as claimed in claim 13, further comprising means for performing the method of any one of claims 2 to 12.
15. An apparatus comprising:

- adaption circuitry configured to adapt statistical parameters of a predefined statistical parametric speech model to create an adapted statistical parametric speech model configured to synthesise speech having a target characteristic. wherein the adaptation of the statistical parameters uses a transform to
- 5 transform speech features derived from speech having a target characteristic.

1/4

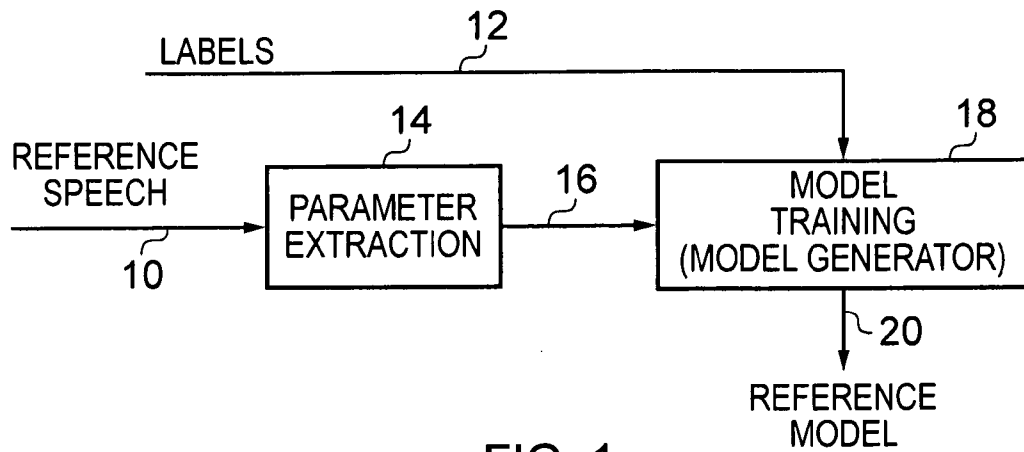


FIG. 1

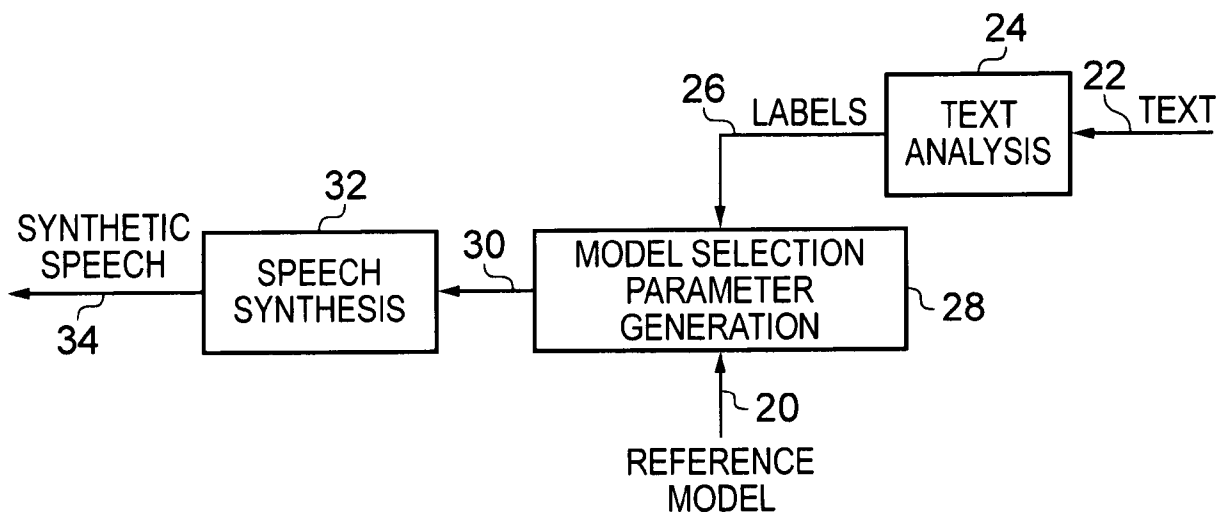


FIG. 2

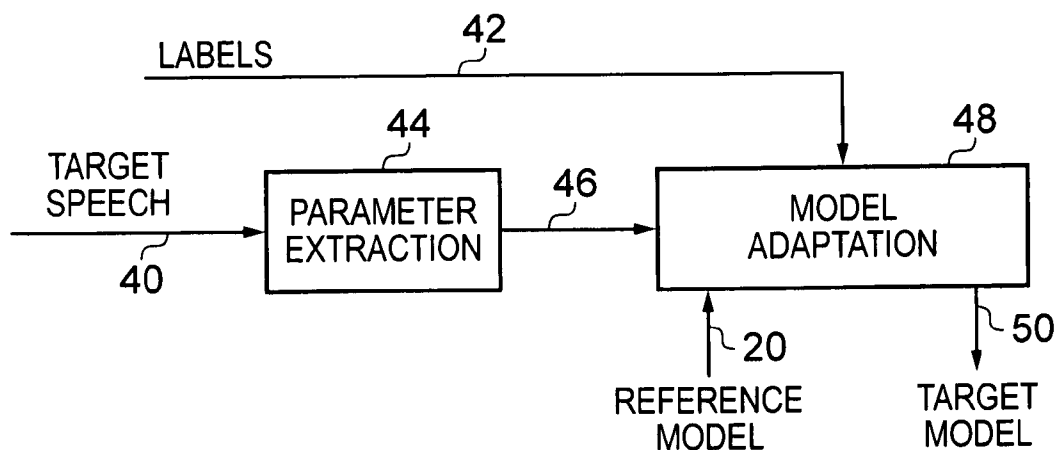


FIG. 3

2/4

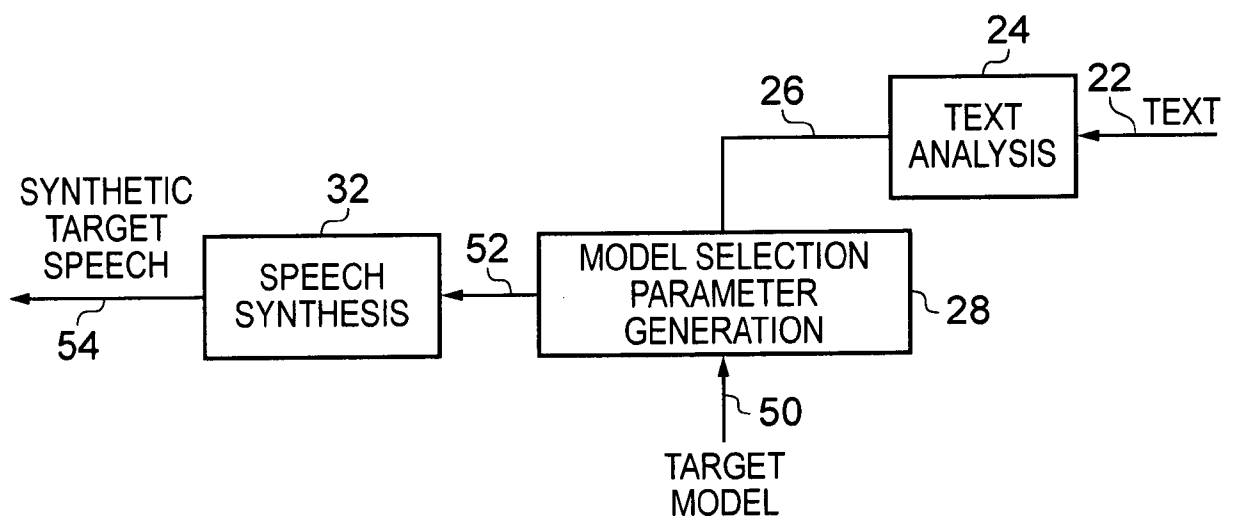


FIG. 4

3/4

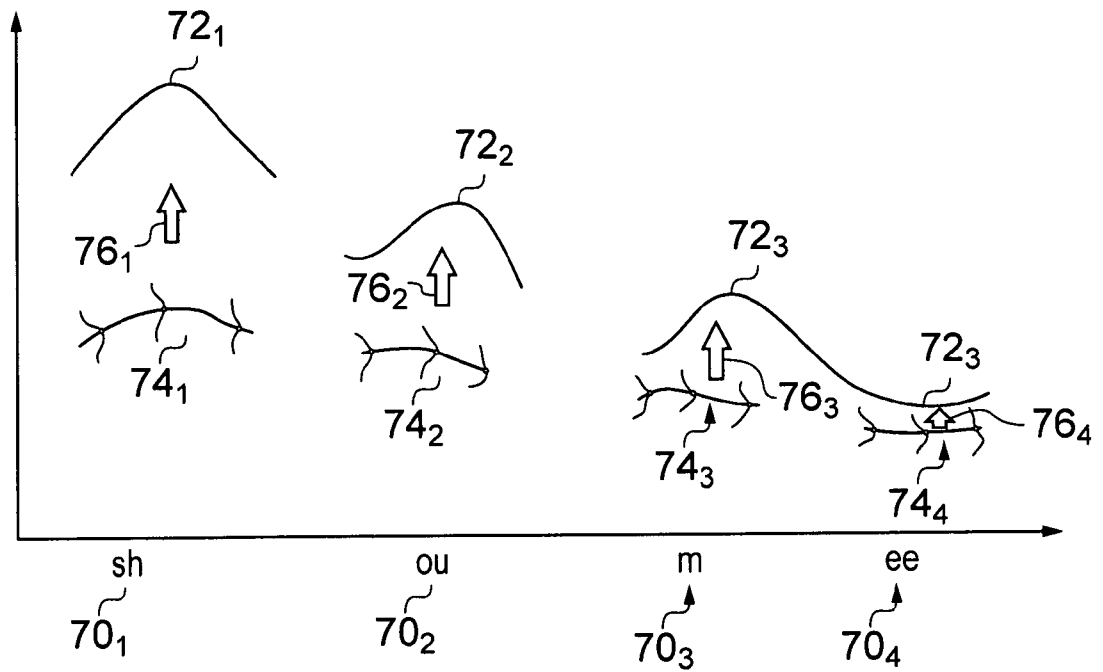


FIG. 6

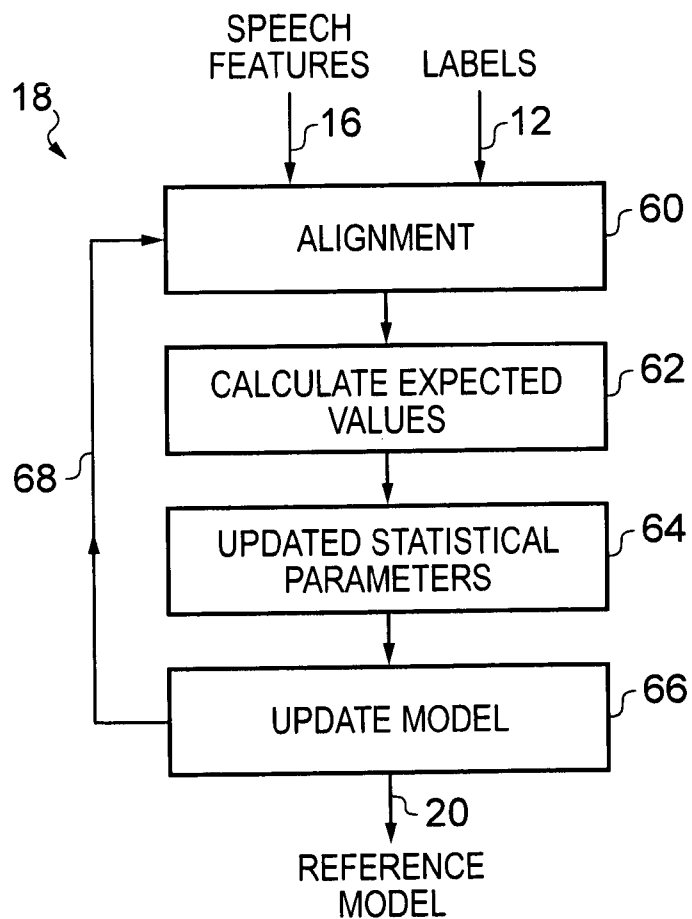


FIG. 5

4/4

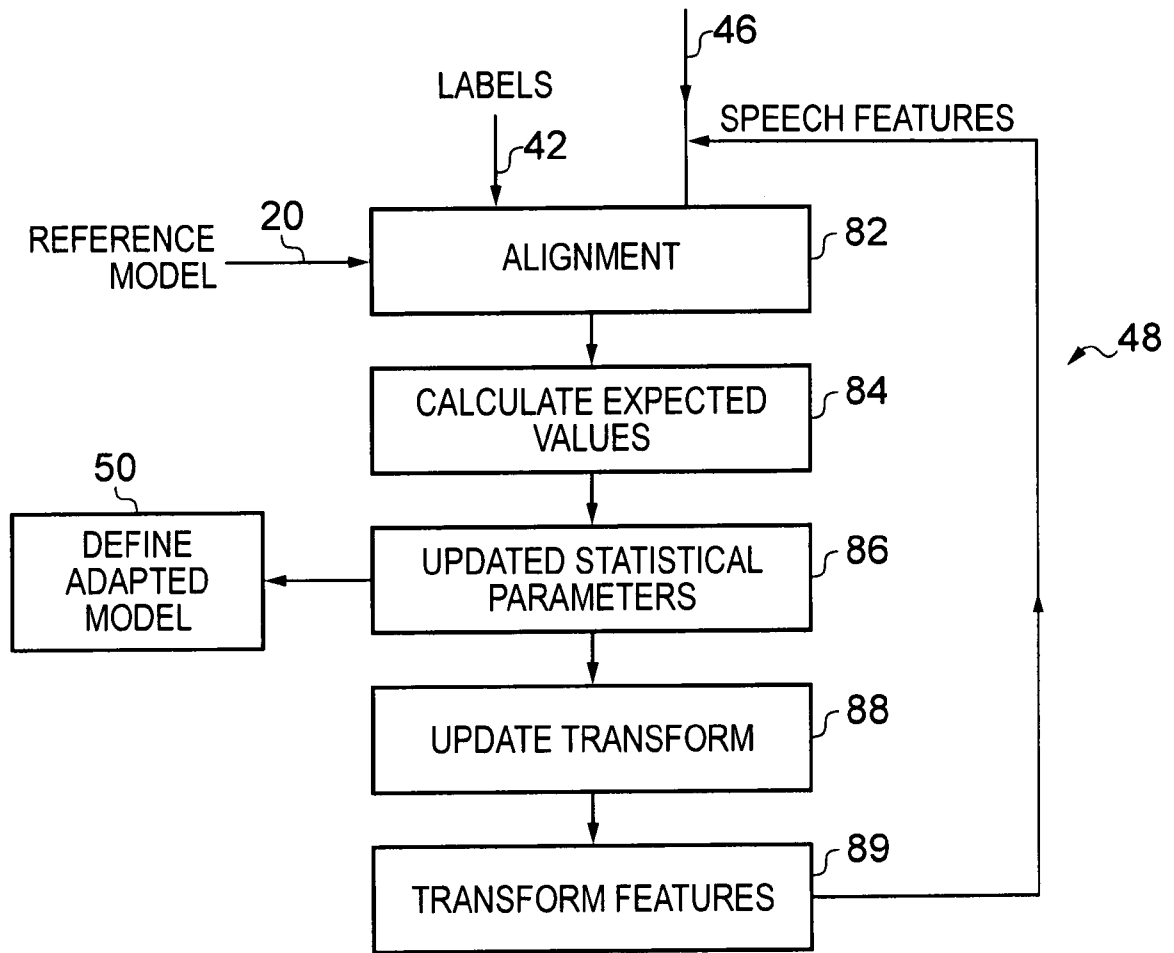


FIG. 7

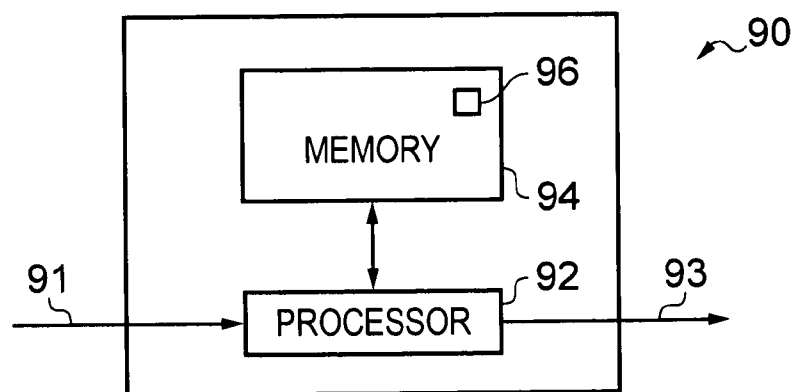


FIG. 8

INTERNATIONAL SEARCH REPORT

International application No

PCT/EP2010/059416

A. CLASSIFICATION OF SUBJECT MATTER

INV. G10L13/02

ADD.

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

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

G10L

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

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

EPO-Internal, INSPEC, WPI Data

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	Junichi Yamagishi, Takao Kobayashi, Steve Renals, Simon King, Heiga Zen, Tomoki Toda, and Keiichi Tokuda: "Improved average-voice-based speech synthesis using gender-mixed modeling and a parameter generation algorithm considering GV" Proc. 6th ISCA Workshop on Speech Synthesis (SSW-6) 1 August 2007 (2007-08-01), XP002605088 Retrieved from the Internet: URL: http://www.cstr.ed.ac.uk/downloads/publications/2007/ssw6-yamagishi.pdf [retrieved on 2010-10-14] the whole document ----- -/-	1-15



Further documents are listed in the continuation of Box C.



See patent family annex.

* Special categories of cited documents:

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier document but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered 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

14 October 2010

Date of mailing of the international search report

25/10/2010

Name and mailing address of the ISA/

European Patent Office, P.B. 5818 Patentlaan 2
NL - 2280 HV Rijswijk
Tel. (+31-70) 340-2040,
Fax: (+31-70) 340-3016

Authorized officer

Quélavoine, Régis

INTERNATIONAL SEARCH REPORT

International application No

PCT/EP2010/059416

C(Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X,P	JUNICHI YAMAGISHI ET AL: "Robust Speaker-Adaptive HMM-Based Text-to-Speech Synthesis" IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, IEEE SERVICE CENTER, NEW YORK, NY, US, vol. 17, no. 6, 1 August 2009 (2009-08-01), pages 1208-1230, XP011263816 ISSN: 1558-7916 the whole document -----	1-15