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

(11) EP 1 640 968 A1

(12) EUROPEAN PATENT APPLICATION

(43) Date of publication:
29.03.2006 Bulletin 2006/13

(51) Int Cl.:
G10L 13/06 *(2006.01)* ***G10L 13/08*** *(2006.01)*

(21) Application number: **04447212.4**

(22) Date of filing: **27.09.2004**

(84) Designated Contracting States:
AT BE BG CH CY CZ DE DK EE ES FI FR GB GR HU IE IT LI LU MC NL PL PT RO SE SI SK TR
Designated Extension States:
AL HR LT LV MK

(71) Applicant: **Multitel ASBL**
7000 Mons (BE)

(72) Inventors:
- **Beaufort, Richard**
  1435 Corbais (BE)
- **Colotte, Vincent**
  54710 Ludres (FR)

(74) Representative: **Van Malderen, Joelle**
pronovem - Office Van Malderen
Avenue Josse Goffin 158
1082 Bruxelles (BE)

(54) Method and device for speech synthesis

(57) The present invention is related to a method to synthesise speech, comprising the steps of

- applying a linguistic analysis to a sentence to be transformed into a speech signal, whereby the analysis yields phonemes to be pronounced and, associated to each phoneme, a list of linguistic features,
- selecting candidate speech units, based on selected linguistic features,
- forming the speech signal by concatenating speech units selected among the candidate speech units.

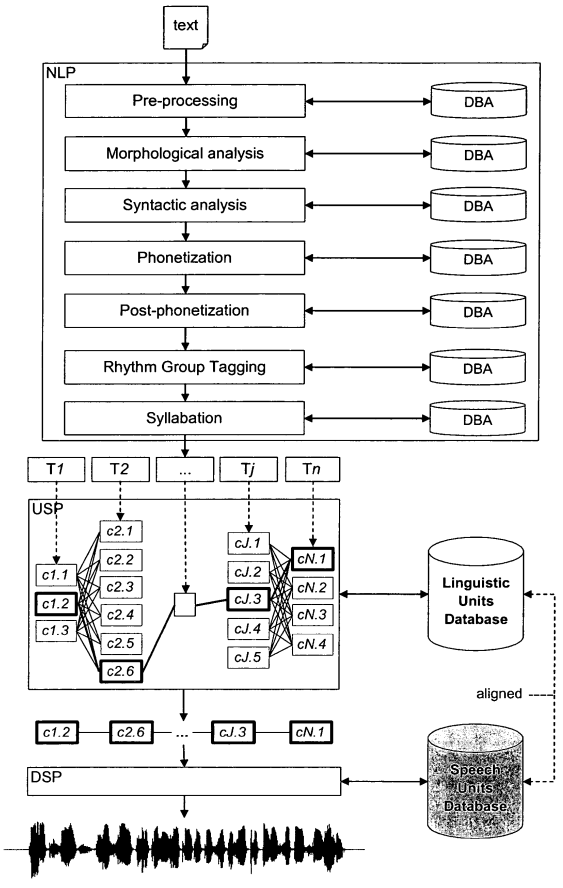

Fig.4

EP 1 640 968 A1

Description

Field of the invention

[0001] The present invention is related to a method and device for speech synthesis.

State of the art

[0002] Nowadays, text-to-speech synthesis systems are based on a sequential and modular architecture, often divided in three major modules: natural language processing, units selection and digital signal processing (see Fig. 1). Natural language processing aims at extracting information that allows reading the text aloud. This information can vary from one system to another but always comprises words, their nature and their phonetisation. Units selection aims at choosing speech units that correspond to the information extracted by natural language processing. Lastly, digital signal processing concatenates the selected speech units and, if needed, changes their acoustic characteristics so that required speech signals are obtained.

[0003] Every synthesis system based on this architecture needs a vocal database containing the different speech units to be used. Most of the time, these units extracted from read-aloud sequences are diphones, i.e. pieces of speech starting from the middle of a phoneme and ending in the middle of the following phoneme (see Fig.2). This means that a diphone extends from the stable part of a phoneme till the stable part of the following phoneme and contains, in its middle part, the coarticulation phase characterising the transition from one phoneme to another, which is very difficult to model mathematically. Using diphones as speech units improves speech generation and makes it easier, because concatenation is performed on their stable parts.

[0004] The first systems using vocal databases for synthesis employed only one sample of each diphone. The underlying idea was to get rid of acoustic variations present in the diphones and dependent from the elocution time: accent, tone, fundamental frequency and duration. In that way, diphones merely are acoustic parameters describing the vocal tract only. Fundamental frequency, prosody and duration have to be regenerated during synthesis. Diphones may need to undergo some acoustic modifications in order to obtain the required prosodic features. This unfortunately leads to a loss of quality: the synthesised voice seems less natural. Besides, despite these modifications, prosody keeps being neutral and listless. Neutral speech units constitute an important drawback to overcome, therefore non-uniform units started to be investigated.

[0005] By *non-uniform* is meant that the speech unit may change in two ways: length and acoustic production. Length variation means that the unit is not exclusively a diphone, but may be either shorter or longer. Longer units imply less frequent concatenation problems. However, in some cases, the corpus constitution (an inconsistency or incompleteness) can impose the use of a smaller unit, like a phoneme or half-phoneme. Therefore a variation in terms of length may be considered in both directions. Variation in terms of acoustic production means that the same unit has to appear several times in the corpus : for the same unit, they may be several representations with different acoustic realisation. By doing so, units are not neutral anymore; they reflect the variations occurring during the elocution.

[0006] Some additional features are provided to the different representations of a same unit such that the system can differentiate them. The art of course lies in the features choice : the features must be relevant and they should not be too few or too many. Relevant means that they provide a good representation of acoustic variations within a unit.

[0007] Every system uses linguistic, acoustic and symbolic features in variable proportions. Linguistic features are directly found by analysing the text, while the choice of both acoustic and symbolic features is based on prosodic models. Among the acoustic features, the fundamental frequency and the duration are the most often used, while the tone is the most recurrent symbolic feature. Each representation of a same unit differs from the other representations by the values of these features.

[0008] The search for speech units corresponding to the units described by natural language analysis often yields several candidates for each target unit. The result of this search is a lattice of possible units, allocated to different positions in the speech signal. Each position corresponds to one unit to be searched for and covers potential candidates found in the corpus (see Fig.3). So the challenge is to determine the *best* sequence of units to be selected in order to generate the speech signal. To do this, the target *cost* and the *concatenation cost* should be used. The *target* cost gives the distance between a target unit and units coming from the corpus. It is computed from the features added to each speech unit. The concatenation cost estimates the acoustic distance between units to be concatenated. The different systems that have been set up determine the concatenation cost between adjacent units in terms of acoustic distance, based on several criteria such as the fundamental frequency, a intensity difference or the spectral distance. Note that said acoustic distance does not compare to the real acoustic distance perceived by a listener.

[0009] The selection of a sequence of units for a particular sentence is cost expensive in terms of CPU time and memory if no efficient optimisation is used. So far, two kinds of optimisation have been investigated.
The first optimisation manages the whole selection. A single unit sequence has to be selected from the lattice. This task

corresponds to finding the best path in a graph. This is usually solved with dynamic programming by means of the well-known Viterbi algorithm.

The second optimisation method consists in assessing the importance of the different features used to determine the target or concatenation cost. Indeed most features may not be considered as equally important: some features affect more the resulting quality than other. Consequently, it has been investigated what would be the ideal weighting for the selection process. The proposed systems however apply a manually implemented weighting, which, as a consequence, is competence based and depends on the operator's expertise rather than on statistic values.

One possible weighting method suggests forming a network between all sounds of the corpus (see *Prosody and Selection* of *source* Units *for Concatenative Synthesis,* Campbell and Black, pp. 279-292, Springer-Verlag,1996). Once this network has been set up, a learning phase can start aiming at improving the *acoustic similarity* between a reference sentence and the signal given by the system. This improvement can be achieved by tuning the features weighting, by successive iterations or by linear regression. This method has two inherent drawbacks: on the one hand, its computational load, still consuming resources even though performed off-line, and on the other hand, the limited amount of features the computation can weight. Most of the time, part of the weighting remains to be done manually. In order to reduce the computational load, one can carry out a *clustering* of sounds to keep only one representative sound, the centroid, on which the selection computation may be performed.

Another weighting method relies on a corpus representation based on a phonetic and phonologic tree (see e.g. *'Non-uniform unit selection and the similarity metric within BT's laureate TTS system',* Breen & Jackson, *ESCA/COCOSDA 3rd Workshop on Speech Synthesis,* pp. 201-206, Jenolan Caves, Australia, Nov. 26-29, 1998). During the selection, they look for candidate units with the same context as the target unit. However, the features they use are not automatically weighted.

[0010] Non-uniform units-based systems try to give synthesised speech a more natural character, closer to human speech than that generated by previous systems. This goal is achieved by using non-neutralised units of variable length. However, the performance of such speech synthesis systems is currently limited by the intrinsic weakness of their prosodic models, restricted to some acoustic and symbolic parameters. These models, corpus- or rule-based, are not sufficient as they do not allow a natural prosodic variation of the synthesised sentences. Yet, the quality of prosody depends directly on how listeners perceive synthesised speech. However, the use of such prosodic models shows a major advantage: the selection of acoustic units that are relatively neutral, limits discontinuities between units to be concatenated further on. As a consequence, spectral smoothing at units boundaries is strongly restricted in order to keep the naturalness of speech units.

[0011] Among the few works that attempt to free themselves from the prosodic model, those of Prudon (see R. Prudon et al., *'A selection/concatenation TTS synthesis system : Databases development, system design, comparative evaluation', ISCA/IEEE 4th Tutorial and Research Workshop on Speech Synthesis, pp. 201-206, Aug.29 - Sept.1, 2001*) make use of only three linguistic features for units selection: the name of the phoneme, its position into the word and its position in the syllable. Unfortunately, units selected by means of these criteria show acoustic discontinuities that require some signal processing. As a result, speech generated by Prudon's system show a less natural character.

Aims of the invention

[0012] The present invention aims to provide a speech synthesis method that does not need any prosodic model and that requires little digital signal processing. It also aims to provide a speech synthesis device, operating according to the disclosed synthesis method.

Summary of the invention

[0013] The present invention relates to a method to synthesise speech, comprising the steps of

- applying a linguistic analysis to a sentence to be transformed into a speech signal, whereby said analysis generates phonemes to be pronounced and, associated to each phoneme, a list of linguistic features,
- selecting candidate speech units, based on selected linguistic features,
- forming the speech signal by concatenating speech units selected among the candidate speech units.

[0014] In a preferred embodiment said selected linguistic features are determined in a training step preceding the above-mentioned steps.

[0015] Advantageously the step of selecting candidate speech units is performed using a database comprising information on phonemes and at least their linguistic features. Preferably the information on the linguistic features comprises a weighting coefficient for each linguistic feature. The weighting coefficients typically result from an automatic weighting procedure. In a preferred embodiment the information is obtained from a step of labelling and segmenting a corpus.

[0016] In another advantageous embodiment the step of selecting candidate speech units comprises the substeps of

- selecting candidate clusters of acoustical representations for each phoneme, and
- computing candidate speech units from the selected candidate clusters.

[0017] Preferably the speech units are diphonic units.

[0018] In a specific embodiment for each candidate cluster a target cost is calculated. Preferably for each candidate speech unit a target cost is calculated from the target costs for the candidate clusters. Typically the concatenation of speech units is performed taking into account said target cost as well as a concatenation cost.

[0019] In a preferred embodiment the linguistic features comprise features from the group {surrounding phonemes, emphasis information, number of syllables, syllables, word location, number of words, rhythm group information}.

[0020] In another object the invention relates to a speech synthesis device comprising a linguistic analysis engine producing phonemes to be pronounced and, associated to each phoneme, a list of linguistic features,

- storage means for storing a database comprising information on phonemes and at least their linguistic features,
- speech units selection means for selecting candidate speech units based on selected linguistic features,
- synthesising means for concatenating speech units selected by said selection means.

[0021] Advantageously the speech synthesis device further comprises calculation means for computing automatically a weighting coefficient for each linguistic feature.

Short description of the drawings

[0022] Fig. 1 represents a Text-to-Speech Synthesiser system.

[0023] Fig. 2 represents the segmentation into phonemes and diphones. "_" corresponds to silence.

[0024] Fig. 3 represents a lattice network for the diphone sequence of the word 'speech'.

[0025] Fig. 4 represents the steps of the method according to the present invention.

Detailed description of the invention

[0026] The present invention discloses a speech units selection system freed from any prosodic model (either acoustic or symbolic) that allows more prosodic variations in synthesised sentences, thereby applying little signal processing at the units' boundaries.

[0027] To synthesise speech without any prosodic model, speech units selection in the method according to the present invention is exclusively based on a features set selected among linguistic information provided by language analysis.

[0028] There are various reasons for this choice. Firstly, any prosodic model, either rules- or corpus-based, relies on a list of linguistic features that allow to choose values for any acoustic or symbolic feature of the model. As a result, a prosodic model is just an acoustic and symbolic synthesis of linguistic features.
Secondly, the prosodic model is deterministic: from a finite list of linguistic features, this model always deduces the same prosodic features. Language however is not deterministic. Indeed, the same speaker could pronounce a given sentence with a single linguistic analysis, in different ways. Parameters having an influence on the pronunciation and prosody of this sentence can be affective or intellective. They can also come from the unconscious and can depend on the elocution time only. These parameters determine the emphasis position, the sounds' duration and the insertion of possible pauses.
Thirdly, variations appearing between several pronunciations of the same sentence are not constraint-free; they have a real influence on the message meaning. However, these constraints can be described by the linguistic analysis. In this way, apart from any affective or intellective emphasis, one may assert that a syllable considered as unstressed may not be emphasised. On the other hand, the emphasis intensity of a stressed syllable can vary a lot. As a function of the linguistic analysis, it is possible to pinpoint parts of sentences where modifications are likely to appear.
These considerations make clear that a sufficiently fine linguistic description allows the management of prosody, without constraining it. The challenge lies in the relevant choice of parameters to be used.

[0029] The synthesis method according to the invention is divided into a training and a run-time phase. In both phases, the same linguistic analysis engine is used for the linguistic features extraction, giving thus some homogeneity to the system. As a first step in the training phase it is necessary to list the relevant linguistic features for selecting the units. Once this list is obtained, the further training consists in a labelling and a segmentation of the corpus as well as a weighting of the linguistic features. Note that in text-to-speech synthesis, a spoken language corpus is always paired with a written corpus that is its transcription. The written corpus helps in choosing labels and features for each unit of the spoken language corpus. It should be noted that the spoken language corpus may as well be called a speech units

corpus or a speech units database. The run-time phase is carried out on a sentence applied to the synthesis system input. First the linguistic sentence is analysed. Then candidate speech units are selected based on selected linguistic features. Lastly, selected units are concatenated in order to form the speech signal corresponding to the sentence. Both phases are now presented in detail.

[0030] The features selection is intrinsically linked to the linguistic analysis engine, the capabilities of which determine the amount of available linguistic information. The exclusive use of linguistic features for selection forces to add supplementary, prosody affecting information to the features typically used (like phonemes around the target, syllabification, number of syllables in the word, location of words in the sentence ...). Very common linguistic features like the phonemes surrounding the target unit and the number of syllables in the word rarely are used in state-of-the-art systems. Consequently, the analysis engine must be powerful enough to determine the required additional information. Said additional information comprises:

- Primary and secondary emphasis of the word, both being strictly linguistic and extractable from phonetisation lexicons,
- Rhythm groups including several types and allowing the implicit determination of the positions where the group emphasis may appear. Rhythm groups also permit to adapt the text syllabification.

[0031] Written and speech units corpora are built separately. By means of the language analysis engine, each sentence of the written corpus is annotated as follows: amount of words and place of the words in the sentence, syllabification and phonetisation of the words, synthesis in terms of articulatory criteria of phonemic contexts for each phoneme. The annotation elements are then discretised as integer values and stored into a linguistic units database wherein each phoneme is linked with its own linguistic features.

The sentences of the spoken language corpus are segmented into phonemes and diphones. All phonemes occurring in the speech units corpus are then collected. For each phoneme the acoustic features useful for the concatenation cost are calculated and also added to the speech units corpus. These acoustic features are the fundamental frequency, LPC (Linear Predictive Coding) coefficients and the intensity. To the phonemes in the linguistic units database additional information is added that allows to pinpoint the speech unit in the_signal: the position of each phoneme (in milliseconds) and the position of its diphonic middle in the speech units corpus.

[0032] The hypothesis is assumed that, because of articulatory differences, each phoneme behaves differently in a same elocution context. Therefore a single weighting for all linguistic features would not be relevant; it is preferable to weight the features independently for each phoneme. For a particular phoneme, one takes all its acoustic representations in the speech units corpus. These representations are split into different clusters, each comprising the acoustic representations to be considered similar. The Kullback-Leibler distance can thereby be used as similarity index.

The optimal number of clusters, varying between 5 (minimum) and 120 (maximum), is automatically computed by maximising the variances ratio. Initially this number is set at 7 clusters of acoustic representations of one phoneme distributed according to their duration d:

$$1.\ d \leq M - 2D$$

$$2.\ M - 2D < d \leq M - D$$

$$3.\ M - D < d \leq M - D/2$$

$$4.\ M - D/2 < d \leq M + D/2$$

$$5.\ M + D/2 < d \leq M + D$$

$$6.\ M + D < d \leq M + 2D$$

$$7.\ d > M + 2D$$

where M denotes the mean duration of all representations for one phoneme, and D represents the standard deviation of this representation.

[0033] Once these clusters are defined, the (fully automatic) linguistic features weighting may start. The objective is to determine to which extent each feature allows to discriminate between several clusters, whereby each cluster is seen as a class to be selected or a decision to be taken. The most appropriate method to do this is by using a *decision* tree. Decision tree building relies on the concept of entropy. Entropy computation for a list of features allows classifying them according to their intrinsic information. The more a feature i reduces the uncertainty about which cluster C to select, the more it is informative and relevant. The entropy of feature i is computed as gain ratio GR(*i*,*C*), i.e. the ratio of Information Gain IG(*i*,*C*) to the Split Information SI(C). The Split Information normalises the Information Gain of a given feature by taking into account the number of different values this feature can take. The weighting coefficient C_i is then computed as :

$$C_i = 2 - \log(1+10GR(i,C)) \qquad \text{(eq.2)}$$

The Gain Ratio allows determining the features ranking between all decision tree levels, and also weights the features during the target cost calculation. The weighting coefficients are also stored in the database.

[0034] At run-time, each time a sentence enters the system, the linguistic analysis generates the corresponding phonemes as well as a list of linguistic features associated to each of them. Every pair {*phoneme, features*} is defined as a target.

[0035] The speech units selection occurs in three steps:

- for each target, pre-selection of phonemic candidates, and target cost calculation for each candidate,
- computation of a diphonic representation of candidate speech units (diphonic units), and
- selection of the speech units minimising the double cost {*target, concatenation*}.

The pre-selection step only keeps phonemic candidates for a given target if they present at least the same label (i.e. the phoneme name) as the target. However, a more drastic pre-selection could restrict the candidates to those that present certain values for some dominant, best weighted features. Let us say, for instance, that right phonemic context for a given target t is the most important, best weighted feature, and that right context has value v. One might choose to keep candidates only if their right context presents value *v* too.

The target cost computation of each candidate phonemic unit is carried out at this stage. In this computation, features are weighted using the weights determined during the training. Target cost *CC* of a candidate *j* for phoneme *i* thus corresponds to a weighted summation of its features:

$$CC(cand_j, pho_i) = \sum_{k=1}^{N} C_k^j \cdot W_k^i \qquad \text{(eq.3)}$$

where:

- ($cand_j$,pho_i) denotes candidate *j* for phoneme *i*,
- N denotes the number of features,
- C_k^j is the value of the feature *k* for candidate *j,* and
- W_k^j is the weight attributed to the feature *k* for phoneme i in the training phase.

[0036] For the diphonic representation, diphonic units to be selected are only those that can be formed from adjacent phonemic candidates in the speech units corpus. However, if a target diphone does not have any candidate, one creates candidates containing the target phoneme partly left or partly right-hand side, according to the diphone needed. The target cost of each diphonic candidate is the sum of the costs of the two phonemic candidates that constitute it:

$$CC(cand_{kl}, dipho_{ij}) = CC(cand_k, pho_i) + CC(cand_l, pho_j) \qquad \text{(eq.4)}$$

where ($cand_{kl}$, $dipho_{ij}$) is the diphone made up of the candidates {*k,l*} selected for the phonemes {*i,j*}.

[0037] Next, the units selection is performed in a traditional way, by solving the lattice with the Viterbi algorithm. In this way the path is selected in the lattice of diphones, which minimises the double cost {*target, concatenation*}. Note that the target cost was already pre-computed at the pre-selection stage, whereas the concatenation cost is determined when running through the lattice.

The concatenation cost has been defined as the acoustic distance between the units to be concatenated. To calculate this distance, the system thus needs acoustic features, taken at the boundaries of the units to be concatenated: fundamental frequency, spectrum, energy and duration. The distance, and thus the cost, is obtained by adding up:

- the fundamental frequency difference,
- the spectral distance (e.g. of *Kullback-Leibler* type),
- the energy difference, and
- the difference in duration of the phoneme that is used as concatenation point. For example, when the system has to concatenate target diphones /pa/ and /aR/, it tries to favour candidate diphones that present more or less the same duration for the half phoneme /a/.

Of course, the sum is weighted. Contrary to that of the target cost however, this weighting is not learned automatically during training: it is manually given, and favours mainly the spectral distance and the difference in fundamental frequency. The double cost {*target, concatenation*} itself is also weighted, such that the target cost and the concatenation cost do not have the same weight in the choice of the best candidates.

[0038] Figure 4 shows a block scheme of a text-to-speech synthesis system that implements the method of the invention. The system is split into three blocks, each corresponding to one of the steps of the run-time phase as described above : the NLP (Natural Language Processing), the USP (Units Selection Processing) and the DSP (Digital Signal Processing). The input to the system is the text that is to be transformed into speech. The output to the system is a speech signal concatenated from non-uniform speech units.

Each block uses databases. The NLP loads linguistic databases (DBA) for each task (pre-processing, morphological analysis,...). The DSP loads the Speech Units Database, from which speech units are selected and concatenated into a speech signal. The USP, in between, loads a Linguistic Units Database, comprising a set of triplets {*phoneme, linguistic features, position*}. The first pair, {*phoneme, linguistic features*}, describes a unit from the Speech Units Database. The last information, *position,* is the position in milliseconds of the unit in the Speech Units Database. It means that both databases describe and store Candidate Units, and are *aligned* thanks to the *position* feature.

The NLP block aims at analysing the input text in order to generate a list of target units (T_1, T_2 ,..., T_n). Each target unit is a pair {*phoneme, linguistic features*}. The second block, USP, works in three steps. First, it selects from the Linguistic Units Database a set of phonemic candidates for each target unit. A target cost computation is performed for each candidate. Candidate diphonic units are then determined together with their target cost and a lattice of weighted diphones is created, one diphone for each pair of adjacent phonemes. Next, it selects by dynamic programming the best path of diphones through the lattice. The DSP block takes selected diphones from the Speech Units Database. Then, it concatenates them acoustically, using a technique of the *OverLap And* Add type: pitch values are used to improve the concatenation.

No signal processing is necessary other than the concatenation itself. Selected units are concatenated without any discontinuity. As a result, linguistic criteria used in the selection prove their relevance.

[0039] The naturalness of generated speech and especially the prosody, make speech synthesis according to present invention suitable for numerous applications, e.g. in public places for information services. The technology can for example be used for advertisement diffusion (broadcasting) in shopping centres. Advertisements of shopping centres must frequently change, which requires frequent and expensive need for professional speakers. The proposed synthesis method only once requires the services of a professional speaker, and subsequently allows pronouncing any written text, without additional cost.

[0040] Another application could be directed to information for travellers in railway stations and airports and the like. Currently, there are a few advertisers who have a perfect control of all languages in which messages have to be stated. As a consequence, the advertiser's accent can reduce the intelligibility of the message. The synthesis system according to the present invention can easily solve this problem.

[0041] Speech synthesis according to the present invention can also generate fluent interactive dialogues. This is related to dialogue systems able to model a conversation and to automatically generate text in order to interact with the user. Two traditional examples are interactive terminals in stations, airports and shopping centres, as well as vocal servers that are accessible by phone. Systems currently used in this context are strongly limited: based on pieces of pre-recorded sentences, they are limited to some basic syntactic structures. Moreover, the result obtained is less natural, because of prosodic discontinuities at words or word-groups boundaries. The synthesis by non-uniform units selection using linguistic criteria is the ideal solution to get rid of these drawbacks, as it is not limited in terms of syntactic structures.

Claims

1. Method to synthesise speech, comprising the steps of

   - applying a linguistic analysis to a sentence to be transformed into a speech signal, said analysis generating phonemes to be pronounced and, associated to each phoneme, a list of linguistic features,
   - selecting candidate speech units, based on selected linguistic features,
   - forming said speech signal by concatenating speech units selected among said candidate speech units.

2. Method to synthesise speech as in claim 1, wherein in a preceding training step said selected linguistic features are determined.

3. Method to synthesise speech as in claim 1 or 2, wherein the step of selecting candidate speech units is performed using a database comprising information on phonemes and at least their linguistic features.

4. Method to synthesise speech as in claim 3, wherein said information on said linguistic features comprises a weighting coefficient for each linguistic feature, said weighting coefficients resulting from an automatic weighting procedure.

5. Method to synthesise speech as in claim 3 or 4, wherein said information is obtained from a step of labelling and segmenting a corpus.

6. Method to synthesise speech as in any of claims 1 to 5, wherein the step of selecting candidate speech units comprises the substeps of

   - selecting candidate clusters of acoustical representations for each phoneme, and
   - computing candidate speech units from said selected candidate clusters.

7. Method as in any of the previous claims, wherein said speech units are diphonic units.

8. Method as in claim 6, wherein for each candidate cluster a target cost is calculated.

9. Method as in claim 8, wherein for each candidate speech unit a target cost is calculated from said target costs for said candidate clusters.

10. Method as in claims 8 or 9, wherein said concatenation of speech units is performed taking into account said target cost as well as a concatenation cost.

11. Method to synthesise speech as in any of claims 1 to 10, wherein said linguistic features comprise features from the group {surrounding phonemes, emphasis information, number of syllables, syllables, word location, number of words, rhythm group information}.

12. Speech synthesis device comprising

   - a linguistic analysis engine producing phonemes to be pronounced and, associated to each phoneme, a list of linguistic features,
   - storage means for storing a database comprising information on phonemes and at least their linguistic features,
   - speech units selection means for selecting candidate speech units based on selected linguistic features,
   - synthesising means for concatenating speech units selected by said selection means.

13. Speech synthesis device as in claim 12, further comprising calculation means for computing automatically a weighting coefficient for each linguistic feature.

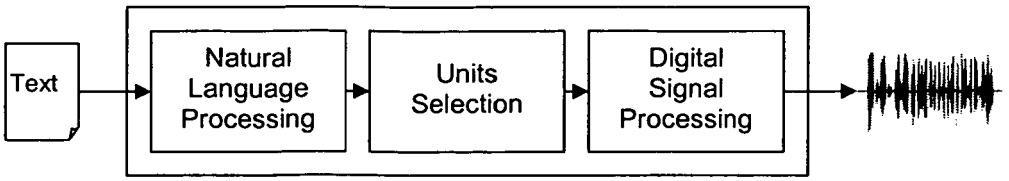

Fig.1

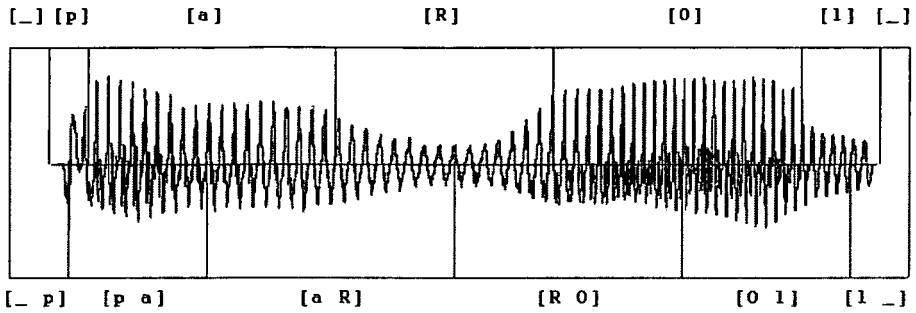

Fig.2

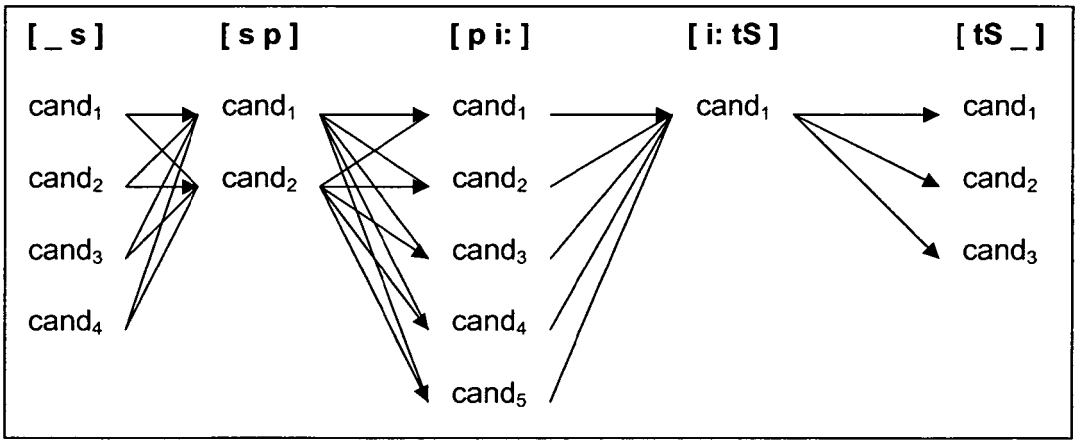

Fig.3

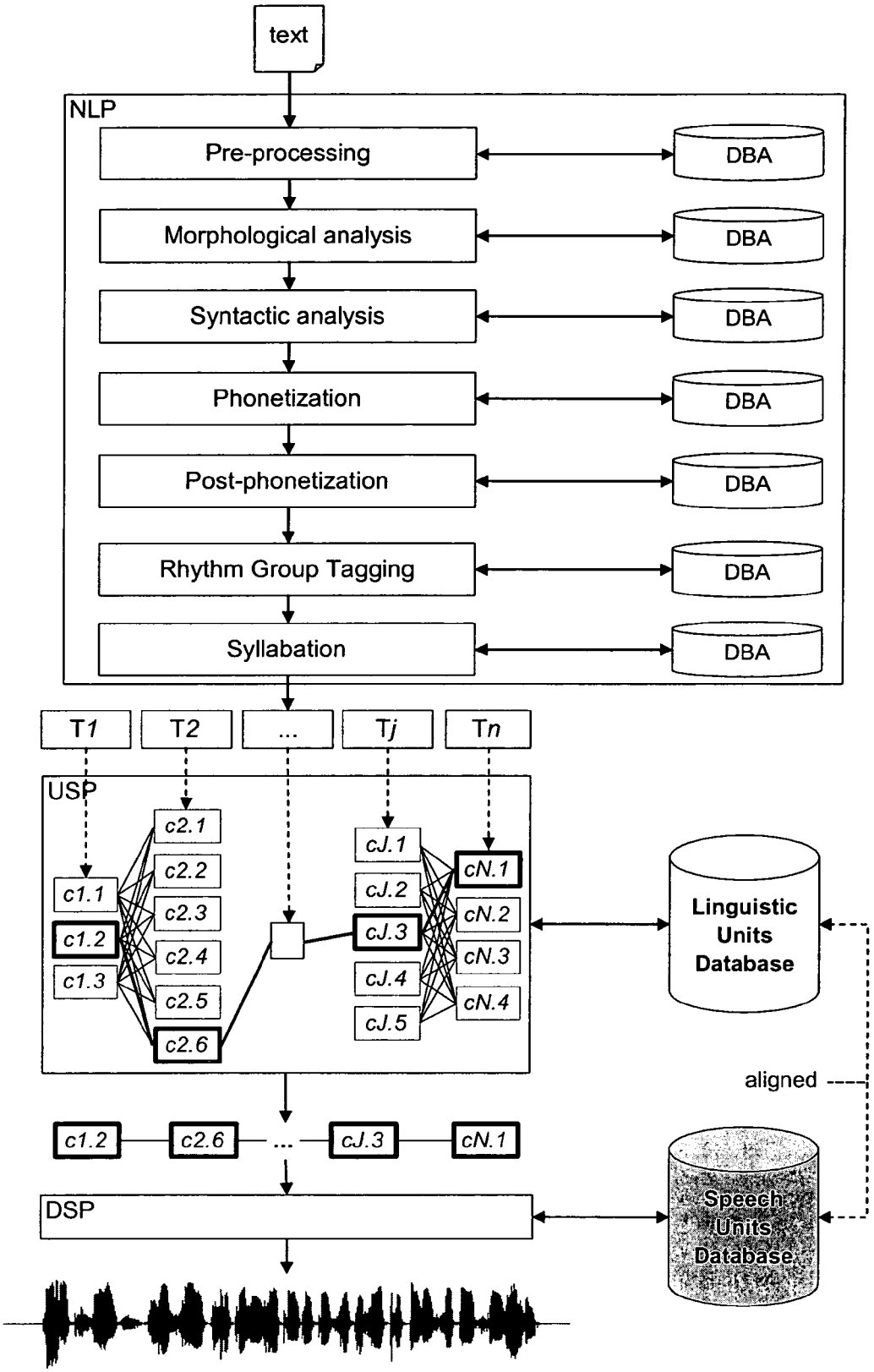
text
NLP
Pre-processing
Morphological analysis
Syntactic analysis
Phonetization
Post-phonetization
Rhythm Group Tagging
Syllabation
DBA
T1
T2
...
Tj
Tn
USP
c1.1
c1.2
c1.3
c2.1
c2.2
c2.3
c2.4
c2.5
c2.6
cJ.1
cJ.2
cJ.3
cJ.4
cJ.5
cN.1
cN.2
cN.3
cN.4
Linguistic Units Database
aligned
c1.2
c2.6
...
cJ.3
cN.1
DSP
Speech Units Database

Fig.4

European Patent Office

EUROPEAN SEARCH REPORT

Application Number
EP 04 44 7212

DOCUMENTS CONSIDERED TO BE RELEVANT

Category	Citation of document with indication, where appropriate, of relevant passages	Relevant to claim	CLASSIFICATION OF THE APPLICATION (Int.Cl.7)
X	COLOTTE V ET AL: "Synthèse vocale par sélection linguistiquement orientée d'unités non-uniformes: LIONS" JOURNÉES D'ETUDE SUR LA PAROLE - JEP '04, [Online] 19 April 2004 (2004-04-19), XP002307516 FEZ, MOROCCO Retrieved from the Internet: URL:http://www.lpl.univ-aix.fr/jep-taln04/proceed/actes/jep2004/Colotte-Beaufort.pdf > [retrieved on 2004-11-25] * the whole document * -----	1-13	G10L13/06 G10L13/08
X	WO 02/097794 A (TAYLOR PAUL ALEXANDER ; AYLETT MATTHEW PETER (GB); FACKRELL JUSTIN WYN) 5 December 2002 (2002-12-05) * page 2, line 11 - line 31 * * page 4, line 4 - line 17 * -----	1,12	
			TECHNICAL FIELDS SEARCHED (Int.Cl.7)
			G10L

The present search report has been drawn up for all claims

Place of search	Date of completion of the search	Examiner
Munich	25 November 2004	Ramos Sánchez, U

CATEGORY OF CITED DOCUMENTS

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

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

& : member of the same patent family, corresponding document

EPO FORM 1503 03.82 (P04C01)

ANNEX TO THE EUROPEAN SEARCH REPORT ON EUROPEAN PATENT APPLICATION NO. EP 04 44 7212

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

25-11-2004

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
WO 02097794 A	05-12-2002	WO 02097794 A1	05-12-2002
		GB 2392361 A	25-02-2004
		US 2004172249 A1	02-09-2004

EPO FORM P0459

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