

- 1 -

## Description

Method for defining a sequence of sound modules for synthesis of a speech signal in a tonal language.

5

The invention relates to a method for defining a sequence of sound modules for synthesis of a speech signal in a tonal language, corresponding to a predetermined sequence of speech modules.

- 10 Automatic methods, carried out by computers, for synthesis of tonal languages, such as Chinese, in particular Mandarin or Thai, normally use sound modules which each represent one syllable, since tonal languages generally have relatively few syllables. These sound modules are concatenated to form a speech signal, in  
15 which process it is necessary to take into account the fact that the significance of the syllables is dependent on the pitch.

- Since these previously proposed methods have a set of sound modules which must include all the syllables in various variants  
20 and contexts, a considerable amount of computation power is required in a computer to carry out this process automatically. This computation power is often not available in mobile telephone applications.

- 25 In applications with a high level of computation power, the previously proposed methods for synthesis of tonal languages have the disadvantage that the given set of syllables does not allow correct synthesis of specific expressions which contain syllables that are not stored in this set, even though sufficient  
30 computation power may be available.

These previously proposed methods have been proven in practice. However, they are not very flexible since they

frequently cannot be adapted to applications where there is little computation power and they do not fully utilize capabilities provided by high computation parallels.

5

A method for language synthesis, which relates to synthesis of European languages, is explained in the thesis "Konkatenative Sprachsynthese mit großen Datenbanken" [Concatenated speech synthesis using large databanks], Martin Holzapfel, TU Dresden, 2000. In this method, individual sounds are stored in their specific left-to-right context as sound modules. Based on "The HTK book, version 2.2" Steve Young, Dan Kershaw, Julian Odell, Dave Ollason, Valtcho Valtchev and Phil Woodland, Entropic Ltd., Cambridge 1999, these sound modules are referred to as triphones. In this sense, triphones are sound modules of an individual phon, although it is necessary to take account of the context of a preceding phon and of a subsequent phon in this case.

20

In this previously proposed method, a group of sound modules (triphones) is stored in a databank for each speech module, which generally comprises one letter. A suitability functions are used to determine suitability distances for sound modules in the respective speech modules, with the suitability distances quantitatively describing the suitability of the respective sound module for representation of the speech module, or of the sequence of the speech modules. The suitability distances can in this case be determined using the following criteria:

25

30

- representativeness of the sound modules;
- manipulation of the sound duration;
- 35 - manipulation of the sound energy;
- manipulation of the fundamental frequency.

- 3 -

When determining the representativeness of the sound modules, a typical spectral centroid of the group of sound modules is defined and a value which is indirectly proportional to the spectral distance between the respective sound module and the centroid is defined as the suitability distance.

When sound modules are concatenated, the fundamental frequency must be manipulated, as a result of which the sound duration and sound energy are also influenced. The corresponding suitability functions are used to determine a measure of the discrepancy from the original state of the sound section as a result of the manipulation.

A corresponding method for determining a sound module which is representative of the speech module is proposed in DE 197 36 465.9. In this document, the suitability functions are referred to as association functions, and the stability distance is referred to as the selection measure. Apart from this, this method corresponds to the method described in the thesis cited above.

The invention provides a method for defining a sequence of sound modules for synthesis of a speech signal in a tonal language, corresponding to a predetermined sequence of speech modules, with this method having a high level of flexibility.

Using the method according to the invention, a sequence of sound modules is defined for synthesis of a speech

signal in a tonal language, corresponding to a predetermined sequence of speech modules, in which

- a group with sound sections, which contains the sound sections which can be associated with the speech module, is chosen corresponding to each of the speech modules in the predetermined sequence,

- a sound module is in each case selected from the respective groups of sound modules for each speech module in that a suitability distance from the predetermined speech module is defined for each of the sound modules in a group on the basis of at least one suitability function, and the individual suitability distances in a predetermined sequence of sound modules are concatenated with one another to form a global suitability distance, with the global suitability distance quantitatively describing the suitability of the respective sequence of sound modules for representation of the respective sequence of speech modules, and with the sequence of sound modules with the best suitability distance being associated with the predetermined sequence of speech modules, characterised in that the sound modules comprise triphones, which each represent only one phoneme with the respective contexts, and the syllables in the tonal language are composed of one or more triphones.

The invention thus provides a method in which the syllables of a tonal language can be composed of triphones. In this case, the principle which is used for synthesis of tonal languages in conventional methods, in which the speech signal is regarded as being composed only of sound modules which describe complete syllables, is not used, and syllables are also composed of triphones. This makes it possible to synthesize syllables very flexibly by means of sound modules.

According to one preferred embodiment, a function which describes the capability to concatenate two adjacent sound modules is used as the suitability function, with the value of this suitability function at syllable boundaries being reduced in comparison to the regions within syllables. This means that the capability to concatenate triphones has a lower weighting at syllable boundaries, so that triphones with a relatively low concatenation capability can be concatenated with one another at syllable boundaries.

According to a further preferred exemplary embodiment, a function which describes the match between the pitch level at the transition from one sound module to an adjacent sound module is used as the suitability function. This results in the pitch level being matched.

The invention will be explained in the following text using the drawings, by way of example, in which:

Figure 1 shows a method for defining a sequence of sound modules for synthesis of a speech signal,

25

Figure 2 shows, schematically, a relationship between partial suitability functions and sound and speech modules,

30 Figures 3-6 each show a partial suitability function in a coordinate system,

Figure 7 shows the profile of the pitch level of two mutually adjacent sound sections, and

35

Figure 8 shows, schematically, the design of an apparatus for speech synthesis.

A text to be synthesized is normally in the form of an electronically legible file. This file contains written characters in a tonal language, such as Mandarin. In a first step S1 (Figure 1), these written characters are converted to the spoken sounds associated with the written characters, with each character in the spoken sounds representing a phoneme or the like.

In a second step S2, a group of sound modules is associated with each phoneme. These sound modules are produced and stored in advance, during a training phase, by segmentation of a sample of speech. Such a sampling of speech can be segmented, for example, by means of fast Viterbi alignment. Each triphone results in a number of suitable sound modules, which are each combined in a group. These groups are then associated with the respective triphones.

A sequence of suitable groups of sound modules is thus determined in step S2, these sound modules being associated with the respective phonemes, with their left-hand and right-hand context. These phonemes with the left-hand and right-hand context are referred to as triphones, and represent the speech modules of the text to be synthesized.

Partial suitability functions, which each result in suitability distances, are calculated in step S3. The suitability distances quantitatively describe the suitability of the respective sound module for representation of the following speech module, or of the sequence of speech modules. Figure 2 shows, schematically, three speech modules SB1, SB2, SB3 to be implemented and three possible sound modules LB1, LB2, LB3. The sound module LB1 is a member of the group which is associated with the speech module SB1. A corresponding situation applies to the pairs SB2, LB2 and SB3, LB3.

The suitability of a sound module for representing a specific speech module may depend on different criteria. In principle, these criteria may be subdivided into two classes. The criteria in the first class govern the suitability of a specific sound module LB1 being able to represent a specific speech module SB1, per se. Since a sequence of speech modules must in each case be converted to a corresponding sequence of sound modules, and sound modules cannot be concatenated with one another in an uncontrolled manner, since undesirable artefacts can occur at the corresponding transitions from one sound module to the other sound module, the second class of criteria represents the suitability of the individual sound modules for concatenation. In this sense, a distinction is drawn between a module target distance between the individual sound modules and the speech modules and a concatenation capability distance between the individual sound modules.

The partial suitability functions are explained in more detail further below.

In step S4, the suitability distances for a sequence of sound modules are linked to form a global suitability distance.

In the exemplary embodiment according to the invention, the value range of all the suitability functions covers the value from 0 to 1, with 1 corresponding to optimum suitability and 0 to minimum suitability. The partial suitability functions can therefore be linked to one another by multiplication using the following formula:

$$E_{\text{global}} = \prod_{\text{Modules}} \prod_{\text{Criteria}} E_{\text{partial}} \quad (1)$$

According to this formula, all the partial suitability distances  $E_{\text{partial}}$  of the individual suitability functions (criteria) for each module are multiplied by one another, and the products which are obtained in the process for each module are in turn multiplied to form the global suitability distance  $E_{\text{global}}$ . The global suitability distance  $E_{\text{global}}$  thus describes the suitability of a sequence of sound modules for representing a sequence of specific speech modules. The value range of the global suitability function is once again in the range from 0 to 1, with 0 corresponding to minimum suitability, and 1 to maximum suitability.

In step S5, that sequence of sound modules is selected which is the most suitable for representing the predetermined sequence of speech modules. In the present exemplary embodiment, this is that sequence of sound modules whose global suitability distance  $E_{\text{global}}$  has the greatest value.

Once the sequence of sound modules which is the most suitable for representing the predetermined sequence of speech modules has been determined, the speech can be produced by successively outputting the sound modules, in which case the sound modules can, of course, be manipulated and modified in a manner known per se.

A number of partial suitability functions are described in more detail in the following text, and these can be used individually or in combination. Figure 3 shows the profile of the partial suitability function  $E_s$  which gives a module target distance as shown in Figure 2, and thus describes the representativeness of the respective sound module for a predetermined speech module. It is thus a measure for the matching of a sound module as a representative, that is to say that a sound module to be selected is a typical, characteristically articulated sound module and is a

suitable representative for the corresponding speech module.

The suitability function  $E_s$  is assumed to be linear  
 5 between the sound section with the "worst" ( $E_s = 1 - S_G$ ) suitability distance and the "best" ( $E_s = 1$ ) suitability distance.

Figure 4 shows, in the form of a suitability function,  
 10 a measure which describes the length manipulation of the respective sound section by the adaptation of a specific fundamental frequency. It is thus a measure of the original duration of the sound section relative to the synthesized duration of the sound section.  
 15 Discrepancies within the range between a lower threshold value  $l_{UG}$  and an upper threshold value  $l_{OG}$  are regarded as not being problematic. Beyond these threshold values, that is to say below the lower threshold value  $l_{UG}$  or above the upper threshold value  
 20  $l_{OG}$ , the partial suitability function  $E_{l\_syn}$  falls exponentially.

This suitability function  $E_{l\_syn}$  is described by the following formula:

25

$$E_{l\_syn}\left(\frac{l-l_0}{l_0}\right) = \begin{cases} \exp\left(-\frac{1}{2} \cdot \left(\frac{l-l_0+l_{UG}}{l_0} \cdot \frac{1}{l_{UG}}\right)^2\right) & \text{for } -l_{UG} > \frac{l-l_0}{l_0} \\ 1 & \text{for } -l_{UG} < \frac{l-l_0}{l_0} < l_{OG} \\ \exp\left(-\frac{1}{2} \cdot \left(\frac{l-l_0-l_{OG}}{l_0} \cdot \frac{1}{l_{OG}}\right)^2\right) & \text{for } l_{OG} < \frac{l-l_0}{l_0} \end{cases} \quad (2)$$

The mean length  $l_0$  is normalized with respect to unity in order to make the discrepancy relative. This partial suitability function  $E_{l\_syn}$  is also normalized with

respect to unity, resulting in a module target distance.

Figure 5 shows a partial suitability function which describes the discrepancy between the pitch level of the sound module and a target fundamental frequency. The pitch level discrepancy relating to a pitch level associated with that sound module in the non-manipulated state should in this case be as small as possible. This partial suitability function  $E_{f\_syn}$  has the following form:

$$E_{f\_syn}\left(\frac{f-f_0}{f_0}\right) = \begin{cases} \exp\left(-\frac{1}{2} \cdot \left(\frac{f-f_0}{f_0} \cdot \frac{1}{f_{OG}}\right)^2\right) & \text{for } o > f-f_0 \\ \exp\left(-\frac{1}{2} \cdot \left(\frac{f-f_0}{f_0} \cdot \frac{1}{f_{UG}}\right)^2\right) & \text{for } o < f-f_0 \end{cases} \quad (3)$$

In this case as well, the frequency  $f$  is normalized with respect to the mid-frequency  $f_0$ . The suitability function  $E_{f\_syn}$  is normalized with respect to unity. An upper frequency parameter is defined as  $f_{OG}$ , and a lower frequency parameter as  $f_{UG}$ .

20

The partial suitability functions shown in Figure 6 describe the discrepancy, which results from the adaptation of a sound section to a fundamental frequency, between the energy in the sound section and a mean value. This partial suitability function is represented by the following formula:

$$E_{E\_ad}(E-E_0) = \begin{cases} \exp\left(-\frac{1}{2} \cdot \left(\frac{E-E_0}{E_{OG} \cdot \sigma_E}\right)^2\right) & \text{for } O > E-E_0 \\ \exp\left(-\frac{1}{2} \cdot \left(\frac{E-E_0}{E_{UG} \cdot \sigma_E}\right)^2\right) & \text{for } O < E-E_0 \end{cases} \quad (4)$$

In this case,  $E_0$  is the mean value (expected value) of the energy  $E$ ,  $E_{UG}$  is a lower energy threshold,  $E_{OG}$  is an upper energy threshold, and  $\sigma_E$  is the energy variance. The suitability function  $E_{E\_al}$  is normalized with respect to unity.

The length  $l$  of the sound section can be used as the criterion instead of the energy. Analogously to Figure 5, this results in a partial suitability function  $E_{l\_al}$  for assessment of the relative discrepancy in the length change of the sound section owing to the adaptation to the fundamental frequency. An upper threshold  $l_{OG}$ , a lower threshold  $l_{UG}$  and a variance for the length  $s_l$  are once again predetermined, so that the suitability function  $E_{l\_al}$  can be represented by the following formula:

$$E_{l\_al}(l - l_0) = \begin{cases} \exp\left(-\frac{1}{2} \cdot \left(\frac{l - l_0}{l_{OG} \cdot \sigma_l}\right)^2\right) & \text{for } l > l - l_0 \\ \exp\left(-\frac{1}{2} \cdot \left(\frac{l - l_0}{l_{UG} \cdot \sigma_l}\right)^2\right) & \text{for } l < l - l_0 \end{cases} \quad (5)$$

The partial suitability functions explained above each result in a module target distance. These suitability functions may be considered individually or in combination for assessment of the sound sections.

The partial suitability function  $E_{f\_syn}$  explained above is used to assess the discrepancy between the fundamental frequency  $f$  of the sound module and a target fundamental frequency  $f_0$ . For synthesis of a tonal language, it is expedient to use a partial suitability function that is modified from this and which assesses the difference between the frequencies of two successive sound sections at their junction point. Figure 7 shows, schematically, the frequency

profile for two successive sound sections LBa and LBb. The sound section LBa ends, and the sound section LBb starts, at time  $t_0$ . There is a frequency difference  $\Delta f$  at this time, since the sound section LBa at the frequency  $f_a$  ends at the time  $t_0$ , at which the sound section LBb at the frequency  $f_b$  starts. In tonal languages, the pitch level is associated with meaning content. The pitch level or frequency of the individual sound sections is thus of fundamental importance for understanding of the synthesized speech. Furthermore, large frequency differences at the transition from one sound section to another sound section result in the formation of artefacts. It is therefore worthwhile assessing the frequency difference between two successive sound sections, with a small frequency difference representing good suitability. A partial suitability function such as this can be formulated, for example, as follows:

$$E_{f\_syn} \left( \frac{f_a - f_b}{(f_a + f_b)/2} \right) = \begin{cases} \exp \left( -\frac{1}{2} \cdot \left( \frac{f_a - f_b}{(f_a + f_b)/2} \cdot \frac{1}{f'_{OG}} \right)^2 \right) & \text{for } o > f_a - f_b \\ \exp \left( -\frac{1}{2} \cdot \left( \frac{f_a - f_b}{(f_a + f_b)/2} \cdot \frac{1}{f'_{UG}} \right)^2 \right) & \text{for } o < f_a - f_b \end{cases} \quad (6)$$

In this case as well, it is once again necessary to provide an upper parameter for the frequency  $f'_{OG}$  and a lower parameter for the frequency  $f'_{UG}$ .

Since this partial suitability function is used to determine a suitability distance between two successive sound modules, this suitability distance represents a concatenation capability distance in the sense of Figure 2.

Further partial suitability functions for describing the concatenation capability of successive sound sections are known from the prior art (see the thesis

"Konkatenative Sprachsynthese mit großen Datenbanken"  
[Concatenated speech synthesis using large databanks],  
Martin Holzapfel, TU Dresden, 2000). The partial  
suitability functions may be used in combination with  
5 the above suitability function  $E_v$ , or else individually,  
in the method according to the invention.

However, for the purposes of the invention, it is  
expedient to weight the suitability functions  $E_v$ , which  
10 describe the concatenation suitability, as a function  
of the region in which the concatenation boundary is  
located. For example, the concatenation suitability  
between two sound sections of a syllable is  
considerably more important than at the syllable  
15 boundary, or at the word or sentence boundary. Since,  
in the present exemplary embodiment, the value range of  
the partial suitability functions is between 0 and 1,  
it is possible to obtain a weighted suitability  
function  $E_{gv}$  by applying a weighting factor to the power  
20 of the unweighted suitability function  $E_v$ :

$$E_{gv} = (E_v)^{g_n} \quad (7)$$

In this case,  $g_n$  is the weighting factor. The higher the  
chosen weighting factor, the more important is the  
25 concatenation suitability between two successive sound  
sections. Suitable values for weighting factors are,  
for example,  $g_1 = 0$  at sentence boundaries,  $g_2 = [2, 5]$   
at word boundaries,  $g_3 = [5, 100]$  at syllable boundaries  
and  $g_4 \gg 1000$  within a syllable. The value of the  
30 concatenation function  $E_v$  thus has a weighting factor  $g_n$   
applied to its power, for which reason small values of  
 $E_v$  with a high weighting factor result in a weighted  
suitability distance close to 0. For the weighting  
factor values stated above, only an unweighted  
35 suitability distance which is only slightly less than

- 14 -

unity can be assessed as being suitable for selection of the corresponding sound sections.

5 The use of such a weighting results in the concatenation of only those sound sections within a syllable which "match" one another very well. Syllables are thus in this way produced by means of individual sound sections or triphones. At syllable boundaries, on the other hand, the unweighted concatenation  
10 suitability may be correspondingly lower as a result of the low weighting. The weighting is once again downgraded somewhat at word boundaries. The use of the weighting factor  $g_1 = 0$  at sentence boundaries means that no concatenation suitability is necessary at  
15 sentence boundaries, that is to say two sound sections whose concatenation suitability distance is equal to 0 may follow one another at sentence boundaries.

20 Figure 8 shows the schematic design of a computer for carrying out the method according to the invention. The computer has a data bus B, to which a CPU and a data memory SP are connected. Furthermore, the bus B is connected to an input/output unit I/O, to which a loudspeaker L, a screen B and a keyboard T are  
25 connected. A program for carrying out the method according to the invention is stored in the data memory SP. Furthermore, a text file which contains the speech modules to be converted to sound modules can be entered in the data memory. The method according to the  
30 invention is then carried out by means of the CPU, with the speech modules being converted to sound modules and being output via the input/output unit on the loudspeaker L. In this case, of course, it is possible for the concatenated sound modules to be modified and  
35 to be altered using normal processing methods.

A feature of the invention is that the tonal language is composed of sound modules which

- 15 -

describe triphones, thus resulting in maximum flexibility. For the purposes of the invention it is, of course, also possible for sound modules also to describe complete syllables in the tonal language. The

5 feature is that sound modules which describe

triphones may also be present, and may be concatenated in an appropriate manner. Particular account is preferably taken of the specific characteristics of a tonal language by the assessment of frequency  
10 differences at transitions from one sound section to a further sound section.

The structures of the tonal language are taken into account in an appropriate manner in the synthesization  
15 process by the weighting, according to the invention, of the suitability functions which describe the concatenation characteristics.

## Patent Claims

1. A method for defining a sequence of sound modules  
for synthesis of a speech signal in a tonal  
5 language in accordance with a predetermined  
sequence of speech modules, in which
  - a group with sound sections, which contains the  
sound sections which can be associated with the  
speech module, is chosen corresponding to each  
10 of the speech modules in the predetermined  
sequence,
  - a sound module is in each case selected from the  
respective groups of sound modules for each  
speech module in that a suitability distance  
15 from the predetermined speech module is defined  
for each of the sound modules in a group on the  
basis of at least one suitability function, and  
the individual suitability distances in a  
predetermined sequence of sound modules are  
20 linked to one another to form a global  
suitability distance, with the global  
suitability distance quantitatively describing  
the suitability of the respective sequence of  
sound modules for representation of the  
25 respective sequence of speech modules, and with  
the sequence of sound modules with the best  
suitability distance being associated with the  
predetermined sequence of speech modules,  
characterized  
30 in that the sound modules are triphones, which  
each comprise only one phoneme with the  
respective contexts, and with syllables in the  
tonal language being composed of one or more  
triphones.
- 35 2. The method as claimed in claim 1,  
characterized

in that a partial suitability distance is in each case calculated using a number of suitability functions for each sound module, and the individual partial suitability distances for the predetermined sequence of sound modules are multiplied by one another to form the global suitability distance.

3. The method as claimed in claim 1 or 2, characterized in that a function which describes the capability to concatenate two adjacent sound modules is used as the suitability function, with the value of this suitability function being weighted differently at syllable boundaries than within syllables.

4. The method as claimed in claim 3, characterized in that the suitability function which describes the concatenation capability is also weighted at word and sentence boundaries.

5. The method as claimed in claim 3 or 4, characterized in that the weighting is carried out by applying a weighting factor (g) to the power of the respective suitability functions.

6. The method as claimed in claim 5, characterized in that the weighting factor ( $g_4$ ) is greater than 1000 within syllables, and the weighting factor ( $g_3$ ) is between 5 and 100 at syllable boundaries.

7. The method as claimed in claim 6, characterized

in that the weighting factor ( $g_2$ ) is between 2 and 5 at word boundaries, and the weighting factor ( $g_1$ ) is equal to 0 at sentence boundaries.

- 5    8.    The method as claimed in one of claims 1 to 7,  
characterized

in that a function which describes the match  
between the pitch levels of two adjacent sound  
modules is used as the suitability function.

10

9.    The method as claimed in one of claims 1 to 8,  
characterized

in that the individual suitability distances in a  
predetermined sequence are linked to one another  
15 by multiplication, with the suitability distances  
being in the value range from 0 to 1, and 1  
corresponding to optimum suitability and 0 to  
minimum suitability.