Speech Synthesis using Vocabulary of Interphoneme Transitions

Almaz Sadenov*
Eurasian National University named L. N. Gumilyov, Scientific-Research Institute, Astana, Kazakhstan; a.sadenov@gmail.com

Abstract

Objectives: This paper presents a modified method of diphone synthesis of Kazakh language. The aim of this work is to create a high-speed synthesizer with acceptable quality. Methods/Statistical Analysis: Diphones represent transitions between sounds (interphoneme transitions). Therefore, to increase the quality of standard method, a technique was used for trimming the edges according to signal quasi-periods and automatic splicing normalization according to the amplitude at the beginning and the ends, which gave a significant increase in quality while maintaining the advantages of this method of synthesis. Application/Improvements: Finally, the described method is implemented as software for speech synthesis of the Kazakh and Russian languages.

Keywords: Diphone, Quasi-Period, Phoneme, Synthesis, Speech

1. Introduction

This paper presents a synthesis method that uses finite dictionary of diphones representing transitions between sounds. Diphone synthesis is a special case of a compiled synthesis known best for solving typical problems of diphone methods for the audibility of diphone splices. Such approach uses diphone methodology of trimming the edges according to their quasi-periodicity and automatic splice normalization according to the amplitude at the beginning and at the end, which results in a major improvement in quality compared with conventional diphone method of synthesis.

As a result, there is a reliable and fast method for the synthesis of speech with fine quality and minimum vocal base. The given method has been exploited in the software of speech synthesis of the Kazakh and Russian languages.

The summary of this paper is presented in the following way. A short description of diphones and methods of its recording in automated regime is given. Then the idea of “quasi-period” is explained with given algorithm for calculation of quasi-periods and proof that is it necessary for calculation. Afterwards, there is a clear application of the normalization procedure as a final stage of such method with a chance of intonational coloring.

2. Existing types of Speech Synthesis

A majority of contemporary speech synthesis methods can be divided into the following types:

- Compilation synthesis
- Synthesis based on rules
- Formant synthesis
- Articulatory synthesis
- Parameterized statistical synthesis – Hidden Markov Model (HMM) - based synthesis.

Each of these types has certain limitations. Being the simplest method, compilation synthesis has audible defects in splicing of audio samples recorded by the narrator. Formant synthesis can be applied without any limitations in real-time systems due to low performance
Speech Synthesis using Vocabulary of Interphoneme Transitions

and poor quality of the received voice. Articulatory synthesis is attributed to robotic voice because of the complexity of building a reliable model of human vocal tract. The most interesting and promising method is HMM-based synthesis.

The biggest disadvantage of the HMM-based generation synthesis approach against the unit selection approach is the quality of synthesized speech. There seems to be three factors, which degrade the quality: vocoder, modeling accuracy, and over-smoothing. The synthesized speech by the HMM-based generation synthesis approach sounds buzzy since it is based on the vocoding technique.

3. Description of Synthesis Method

3.1 Fixed Areas

Diphone model assumes that some fixed areas can be singled out from speech whose sound is not affected by the neighboring sounds. Amidst of these fixed areas there is a boundary between diphones. However, the total amount of diphones in a specific language is not lower than the total number of allophones in the same language.

This fixed area shall be referred as medium of the sound phoneme.

This same rule applies to diphones with unvoiced phonemes as well. It is important to consider combinations of phonemes with innate sound while compiling a phoneme database whereas some phonemes are not pronounced at all.

The number of diphones depends on the phonotactics of the language. For instance, in Spanish there are about 800 diphones whereas German language has approximately 2500. For Russian, this value reaches almost 8000.

Such method of diphones separation permits to splice diphones in places of maximum speech matching, which increases the quality of connection. The issues that shall be look at are the amplitude difference at the ends of diphones and audible “clicks” (spectral gaps) which occur while even splicing identical phonemes.

3.2 Record of Diphones

As a basis, an 8-bit digitization of sound signal is used with a sampling frequency of 22,050 Hz so that its values have \(2^8=256\) gradations: from 0 to 255.

The system is assumed to be used in the laboratory conditions having no significant external noise. While setting up the recording system to the moment of pushing the corresponding button 30,000 samples of “silence” are recorded. Afterwards, successive segments of the recorded signals are analyzed with 300 samples at a time. For each of segment the ratio is calculated

\[
\frac{V}{C},
\]

where

\[
V = \sum_{i=0}^{259} |x_{i+1} - x_i|
\]

V-Digital equivalent to the total variation, C - the number of consistency points, i.e. the moments of time for which the next moment of the signal remains unchanged. The value of (1) is automatically determined, which is typical to used sound card as the most frequently occurring value in the array. It increases by 0.1 and is recorded into the control file of the synthesis system as the starting threshold and the result is multiplied by 2 times as the final threshold.

To automate the recording of diphones, speech coming from a microphone can be processed in successive segments by 300 samples value (1). As soon as this value exceeds the starting threshold no less than five times the following position is saved. From this position, samples are added to the “buffer 1” until the specific moment. This moment is defined by the value (2) at this and successive 10 000 samples less than the last threshold. After this recording is stopped. Since exceeding in the starting threshold may happen not due to the emergence of speech but due to random noise, the content of “buffer 1” is analyzed for the presence of speech by calculating the sequence of quasi-periods. If there are at least five consecutive quasi-periods, the value of which exceeds a predetermined threshold value, the content of the “buffer 1” is considered the speech and is sent to the “buffer 2”. Thus, the speaker can utter the necessary list of diphones continuously without the need for interaction with the elements of the interface, which makes it easy to fill up the diphones base in a quick way.

3.3 Quasi-periods

There is a belief that there are difficulties in trying to use the diphone database for speech synthesis. The fact is that if the speech is formed by connecting the diphones, then
the splicing points show noticeable differences. Resulting breaks are obvious aurally – the speech, “spliced” from separate diphones sounds unnatural.

To resolve this issue we should refer to the nature of the sound. As it is known, the vowels and voiced consonants are generated by the airflow and the vocal cords make vibrations similar to periodical ones.

Therefore, the corresponding signals (amplitude-time representations) are quasi-periodic (see the exact definition of a quasi-period below). Here is, for example, the amplitude-time representation of sound “A”.

Figure 1. The result of automatic segmenting the voice signal phase into quasi-periods.

In this Figure 1, the vertical marks (boundaries of quasi-periods) are placed automatically according to the following algorithm.

- The value is calculated

\[ L_k := \sum_{i=0}^{k-1} x_{n_i+i} - x_{n_i+k} \]  

(3)

where \( n_0 \) - sample number from which the current quasi-period has started in the buffer,

\[ MIN \leq k \leq MAX, \]

- \( k = k_0 \) is determined, in which the value \( t_k \) takes minimum value. By definition \( t_k \) is a length of quasi-period. \( MIN \) and \( MAX \) numbers are determined by the height of speaker’s voice. For tenor we can take

- \( MIN = 60, \quad MAX = 200. \)
- By taking the end of identified quasi-period for the beginning of the next quasi-period we identify the second quasi-period and the same way for others.

Segments of the signal corresponding to hushing sounds or pause-like segments corresponding to unvoiced plosives and sound like “F” do not have a quasi-periodic structure. Therefore, when applying described algorithm to these segments the value of (1) will be reduced not due to closeness to zero difference standing under the sum and due to minimizing the number of summands. As a result, the numbers which in the case of voice sound provide the values of quasi-periods will be close to the value of \( MIN \).

By the way, it allows for distinguishing quasi-periodic areas from non-quasi-periodic ones and for example, using this difference to analyze the presence of the speech signal.

Thus, the compilation of diphones base we can automatically trim the edges of an audio signal according to the quasi-periods, i.e. obtained diphone starts with the first quasi-period and ends on the last quasi-period. For instance, for the sound “A” the result will look like as shown in Figure 2.

Figure 2. Result of automatic trimming of the phoneme “A” according to the quasi-periods.

Accordingly, we can splice different variations of interphoneme transitions with the phoneme “A”, for example “ma” + “am” and so on without discontinuities in diphones splicing points. After the procedure of trimming according to the quasi-periods, move to the normalization of diphones.

### 3.4 Normalization

The problem to be solved is the difference of amplitude differences at the end and beginning of two different diphones. This means that the end of the first diphone may have greater amplitude than the beginning of the following diphone when splicing and vice versa. Accordingly, in the synthesis of the two diphones it is necessary to calculate the value of signal amplitude at the edges for normalization procedure.

Let us consider the example of splicing two diphones “mo” and “ol”. We need to splice together “mo”+“ol” to get the beginning of the word “moloko” (we defined that for the given word we segment the following diphones “m0”-“mo”-“ol”-“lo”-“ok”-“ko”-“o2”)

Let us imagine recorded diphones as the sum of the signal values at a point of time
Speech Synthesis using Vocabulary of Interphoneme Transitions

$$V = \sum_{i=0}^{n} |x_i|$$  \hspace{1cm} (4)

where $x_i$ - the value at the moment of time.

Accordingly, the diphone “mo” will be represented as $V_1$

$$V_1 = \sum_{i=0}^{n} |x_i|$$  \hspace{1cm} (5)

And the diphone “ol” as $V_2$

$$V_2 = \sum_{i=0}^{n} |x_i|$$  \hspace{1cm} (6)

Further when determining the value $|x_i|$ at a point of time as a set $M_1$ (values from $n/2$ up to $n$ for $V_1$) and $M_2$ (values from the beginning of 0 up to $n/2$ for $V_2$).

$$M_1 = \{x_i | i \in 1..N\}$$

$$M_2 = \{x_i | i \in 1..N\}$$

we define the amplitude $A$, as the difference between the maximum and the minimum $|x_i|$ in the interval $n$.

$$A = \max M - \min M$$

Then we can calculate the amplitude $A_1$ for the first $V_1$ and the amplitude $A_2$ for $V_2$.

$$A_1 = \max M_1 - \min M_1$$

The amplitude of the second diphone accordingly is:

$$A_2 = \max M_2 - \min M_2$$

And calculate the average value of ratio $k$ for normalization.

$$k = \left\| \frac{A_2}{A_1} \right\|$$

Using this ratio, we can normalize $V_2$ (diphone “ol”) with respect to $V_1$ (diphone “mo”) as follows:

$$V_2 = \sum_{i=0}^{n} k |x_i|$$  \hspace{1cm} (7)

To sum up, applying the normalization of diphones from the first to the last one we get a normalized synthesized sound, which is ready for emotional coloring.

3.5 Intonation

The most important key for defining intonational boundaries while voicing a written text are punctuation marks. However, there is no one-to-one correspondence. Generally, to determine the boundaries of intonation except punctuation keys we need information on the boundaries of the main components of the syntactic sentence constituents requiring either a full syntactic analysis of a phrase, or the use of probabilistic syntactic and intonational heuristics. Both require special study and are specific case of many text-to-speech-systems including technologically well advanced ones (e.g. text-to-speech-system for the English language)\(^4\).

Minimum necessary intonation characteristics are pausing and stress on the necessary syllable. For the simplest case of intonation we can use one of the algorithms for the tone shift according to the outline of the word\(^5\).

4. Conclusion and Perspectives

In this paper, we did not describe the algorithm of text transcription that takes the input text into the required set of diphone transcriptions, as it is the task for text processing and is related to the synthesis indirectly. The issue should be seriously considered for further work on the intonation of the synthesized text.

The main purpose of the study was to provide a simple method for the synthesis of acceptable quality, which could be used in devices with limited hardware capabilities while bringing quality to the required maximum.

With the help of the described method in this paper a working program was developed which synthesizes the Russian and Kazakh speech. In addition, the following tools were also created:

- Diphones recording tool that automatically determines the beginning and end of words. Saved signal is cut off at the edges according to the quasi-periods; amplitude values of the beginning and the end of diphone are also saved.
- Tool for making diphones dictionary. It automatically generates all the unique combinations of interphoneme transitions from the input word list.
- Speech synthesis tool using the methodology described above.
• Further work in this direction is to improve the intonation of speech via adding the necessary characteristics to diphones transcriptions. However, even in the current implementation the quality is sufficient for use in human-computer interaction and mobile devices.

5. References

1. Frolov AV, Frolov GV. Synthesis and speech recognition. Contemporary Solutions. 2003.

2. Speech Synthesis. 2013. Available from: http://en.wikipedia.org/wiki/Speech_synthesis

3. Shelepow VY. Lectures about speech recognition. IPSHI; Donetsk. 2009. p. 14-15.

4. Vintsyuk TK. Analysis, recognition and interpretation of speech signals. Kiev. 1987.

5. Krivnova OF, Zakharov LM, Zinoviev NV, Strokin GS, Babkin AV. Experience in the development of automatic speech synthesis for the Russian language. Proceedings of the 9th session of RAS, AKIN; Russia. 1999.