The Quality Evaluation of Allophone Database for English Concatenative Speech Synthesis

Karina Evgrafova.

Department of Phonetics
Saint-Petersburg State University, Saint-Petersburg, Russia
Karinaevgr@mail.ru

Abstract
In the paper the procedure of the sound unit inventory construction for English concatenative speech synthesis is described and the results of perceptual tests aimed at evaluating the quality of allophonic inventory are presented. The main criteria of the evaluation were the degree of naturalness and intelligibility of the resulting synthesized speech.

1. Introduction
The paper deals with the sound database formation for English concatenative speech synthesis. The fundamental principles of the English allophonic database formation have been already described by the author [2].

In the present paper, the procedure of the main steps the sound unit inventory construction is described and its results are discussed. The research has been carried out in three steps. In the first step, the acoustic material for the allophonic database was obtained. As a second step, the optimization of the allophonic base was performed. And, finally, perceptual tests were conducted to evaluate the quality of the allophonic inventory developed.

2. Sound Material Preparation
In creating the database of sound elements it was first necessary to obtain basic material, containing all segments required for the organization of a word form. In the allophone-based model these segments are allophones, that is realizations of English phonemes in definite phonetic contexts.

Descriptions of all combinatorial positional allophones of English phonemes can be found in literature [3], [4]. Taking into account these descriptions, a lexicon of 3300 words and phrases containing English allophones in specified contexts was formed [2]. The lexicon was recorded from a female speaker. The recorded speech was digitized at F16KHz.

In the following step, segmentation of the recorded speech was performed. Sound units corresponding to the realizations of allophones in the natural physical limits were extracted from the natural speech stream. It should be noted that in the allophonic model the precise definition of the boundary between two neighboring allophones is of great importance [8]. The quality of the constructed allophone depends on how accurately the boundaries are determined. It is necessary for the allophone to preserve the information contained in transitions, which provides the naturalness of the concatenation. To achieve this aim all coarticulation rules must be taken into account [3].

Thus a lot of effort was invested in the definition of the physical boundaries between allophones during the segmentation process.

To evaluate the accuracy of the segmentation two types of tests were carried out [5]. An extracted from a word allophone was inserted into a word or a syllable having a similar phonetic context. (E.g. /i/ from wick was inserted into wig instead of /i/ from wig). If it led to perceptible change of acoustic properties of the word the right or the left boundary was shifted and the listening experiment was repeated. The physical boundaries of the allophone were set finally only after the needed perceptual effect was achieved.

By means of the EDS program (a program for digital treatment of speech signals developed at St.-Petersburg University of Telecommunications) the extracted allophones were combined into a sound file. After that the listening experiment was conducted to evaluate naturalness of the string synthesized. After testing all extracted allophones were saved in separate sound files.

To improve the quality of the synthesized speech certain consonant clusters were included into the database as one unit. It concerned the following clusters:
1. plosive+fricative in a final position (ps, bz, ts, dz, ks, gz);
2. plosive+lateral sonant /l/ (/tl, dl, bl, kl, gl/);
3. plosive+nasal sonant /n/, /m/, /ng/ (bn, pn, tn, dn etc.);
4. 4. plosive+ plosive/affricate (td, gd, kt, pt etc.).

Thus, as the result of segmentation, the allophonic database inventory was obtained.
At that stage it contained about 3000 concatenation units. As it is well-known, any sound database must be compact. So it was necessary to include all the units needed for quality during synthesis and at the same time to minimize the size of the inventory. Consequently, the optimal proportion between the number of the database elements and the quality of the synthesized speech should have been found. This task was accomplished in the following step.

3. Optimization of the Base of Allophones

After the database of sound elements had been developed, there arose a possibility to reduce the number of the database elements. The previously defined set of all the theoretically necessary allophones was reduced by enlarging the classes of allophone contexts and looking for basic allophones which were still more tolerant to specific contexts. As a result of numerous experiments, which were carried out under a thorough audio control, a sufficient set of fewer basic allophones was defined. The set included all the necessary basic allophones which could be used instead of certain combinatorial allophones [7]. (E.g. in the course of listening experiments it was found out that vowel allophones with /u/ in the left-hand context are similar considerably to those having /w/ in the left. The choice was made in favour of the allophones following /w/ as the combination /w/+vowel is more frequent than /u/+vowel).

Thus the number of the left-hand contexts for vowels was reduced up to 11 classes and the right-hand contexts up to 9 classes. The consonant contexts are presented by the 5 right-hand and 10 left-hand classes (in more detail the results of the optimisation are presented in [3]).

Another way to reduce the number of allophones in the database was used. All the allophones that are possible at word boundaries only were excluded from the database. A micro pause which appears between words as a result of this reduction does not affect the naturalness of the synthesized speech but at the same time improves its intelligibility. Thus it was possible to reduce the size of the allophonic database up to 1200 elements (1000 vowel allophones and 200 consonant allophones).

As a result of the optimization, important theoretical and practical results were obtained. The conducted experiments showed that a number of vowels with different right-hand or left-hand contexts having different articulatory characteristics do not differ acoustically. Consequently, the number of basic allophones which are different from the acoustical point of view turned out to be significantly less than that of the traditionally defined articulatory allophones described in literature [3], [4].

4. Testing the quality of the database inventory

It is well known that major challenges of concatenative speech synthesis are naturalness and intelligibility of the resulting synthesized speech. A common and annoying artifact of concatenative synthesis is an audible discontinuity at concatenation join points. Concatenative speech synthesis quality depends partly on the minimization of audible discontinuities of sound units taken from different contexts. To detect and eliminate possible discontinuities perceptual experiments should be conducted and the results obtained should be given a comprehensive phonetic analysis.

4.1. Methods

4.1.1. Material

A set of 160 isolated words was synthesized using the sound units of the acoustic inventory which had been earlier developed. The test utterances had different syllabic structure. The set included both monosyllabic and polysyllabic (up to 6 syllables) words. Overall, synthesized utterances covered 360 most frequent vowel and consonant allophones of English phonemes.

A very simple concatenation method was used to synthesize the test stimuli. Required allophones were selected from the database inventory by means of a special program and then were concatenated in a separate sound file. No prosodic modification was applied after selection to sound units being concatenated.

4.1.2. Listeners

Fourteen adult volunteer listeners participated in perceptual tests. They represented diverse language backgrounds. One group of listeners included 3 native speakers of English. The other group of listeners consisted of 11 Russians: 5 students of English and 6 teachers of English working at the Department of English Language and Literature of Saint-Petersburg University. The tests performed by the two groups of listeners were conducted independently but they followed the same procedure.

4.1.3. Procedure

The perceptual tests were conducted in a multi-media classroom at the Phonetics Department of Saint-Petersburg University using the relatively high quality audio equipment available there. The volume was adjusted to suit their individual preferences. Stimuli
were digitized at 16 kHz. Listeners initiated the presentation of each stimulus by clicking an icon. Written instructions to listeners were provided at the beginning of a test. They included a set of tasks and a number of response types for each stimulus example. The responses were made by filling in a special questionnaire presented in the form of a table.

Table 1. The Listener’s Questionnaire (a fragment).

<table>
<thead>
<tr>
<th>№</th>
<th>Intelligibility</th>
<th>Naturalness</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>comprehensible</td>
<td>acceptable</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>incomprehensible</td>
<td>unacceptable</td>
<td></td>
</tr>
</tbody>
</table>

After hearing a test stimulus, a listener reported whether or not (s)he found the stimulus word comprehensible selecting one of the responses in the column Intelligibility and scored its naturalness on a scale of 1-2 (acceptable/unacceptable) selecting a response in the column Naturalness. Listeners were also encouraged to justify their judgements in the column Comment. It should be noted that using a relative score instead of an absolute one is found by some researchers to be more preferable than to ask listeners to state the quality of a test stimulus on some absolute scale [1].

Each stimulus was heard twice by a listener. The entire test took approximately 30 minutes to complete.

4.2. The Results of Perceptual Tests

There were many similarities between the results obtained from the two groups of listeners. Therefore both kinds of results were analyzed together. Pooling all the listeners’ responses showed that the intelligibility rate was 100%. All synthesized utterances presented in perceptual tests were found fully comprehensible. The naturalness evaluation, however, did not yield so consistent results. The table below shows naturalness rate.

Table 2. Naturalness rate.

<table>
<thead>
<tr>
<th>Types of responses</th>
<th>Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>acceptable</td>
<td>85%</td>
</tr>
<tr>
<td>unacceptable</td>
<td>15%</td>
</tr>
</tbody>
</table>

As it is shown in Table 2, 85% of the synthesized utterances were evaluated acceptable. All acceptable examples were commented on by as nearly indistinguishable from natural utterances. A lower rate of naturalness, however, was observed for 15% of test utterances.

To improve the quality of the utterances that had been given a low evaluation, a comprehensive phonetic analysis of concatenation "errors" was made. Its aim was to indicate if perceptible concatenation "errors" were caused by spectral discontinuities at the junctions of sound units or determined by prosodic features of a synthesized signal. The phonetic analysis detected a number of features of synthesized signals which detracted from the naturalness of the test utterances evaluated as unacceptable.

1. Inappropriate intensity of some unstressed vowel allophones. It should be caused by the fact that they were originally cut from a stressed syllable where the intensity of a vowel normally exceeds that of an unstressed vowel.

2. Inadequate duration of some vowels. It is known that the duration of English vowels in stressed syllables is determined by a following consonant: vowels sound longer when followed by a voiced consonant while a voiceless consonant noticeably reduces a vowel. For instance, /i:/ in bead /bi:d/ is significantly longer than /i:/ in beat /bi:t/ [4], [5]. Thus inadequate duration of some vowels followed by a voiced consonant in the test utterances can be explained by the fact that they originally had a voiceless consonant in the right-hand context.

3. Inadequate duration of some allophones of sonorants. The duration of sonorants depends on a following consonant in the same way as that of vowels. A voiceless consonant in the right-hand context to a great extent reduces a sonorant; when followed by a voiced consonant it sounds long. E.g. /n/ in spend is longer then in spent. In test utterances some sonorants sounded shorter then it was required by a right-hand context.

4. Lack of coordination in pitch characteristics of concatenated sound units. This mismatch can be explained by the fact that fundamental frequency of the units selected to build up an utterance to be synthesized is usually different from the one requested by the prosodic model of the utterance. Matching of waveform elements originally cut from different parts of the speech corpus is of great importance to the quality of the synthesized speech.

The information obtained after analyzing concatenation "errors" was used to improve the naturalness of low-evaluated test utterances. The amplitude of unstressed vowels was damped and the duration of some vowels and sonorants was manually altered according to the requirements mentioned above.

To eliminate the mismatch of pitch characteristics, a fairly simple smoothing procedure available in the Wave Assistant program (a program for digital
treatment of speech signals developed at St.-Petersburg University of Telecommunications) was implemented. After that one more listening test, which was performed by the same listeners joined by 2 new ones, was conducted to mark the progress in the quality of "corrected" utterances. As a result, all of the test utterances were found to be acceptable. Thus the conducted perceptual tests showed that perceptible concatenation "errors" were mainly caused by prosodic features of a synthesized signal. Lack of naturalness in the “unacceptable” test utterances was not determined by spectral discontinuities of units at the concatenation points. This fact confirmed the accuracy of physical boundaries definition between the segments and also demonstrated that the database inventory had been formed in a correct way. As to prosodic characteristics of a sound unit such as duration, intensity and pitch, in a complete text-to-speech system they can be determined by a specially formed set of rules which should be applied to each particular utterance being synthesized. Alternatively, special techniques of prosody modifications can be incorporated to produce speech with the expected prosodic characteristics. These modifications, however, can result in distortion of individual characteristics of speaker's voice and in lowering of naturalness of synthesized speech.

5. Discussion

The present research was aimed at solving both theoretical and applied problems. Theoretically, it was necessary to consider the realization of each of English phonemes in all possible phonetic contexts in order to define a set of basic allophones that could be used instead of certain combinatorial allophones. The obtained set of basic allophones includes only acoustical allophones, that is allophones different from the acoustical point of view. The sounds that have different right-hand or left-hand contexts and different articulatory characteristics but at the same time are identical acoustically were excluded from the set of basic acoustical allophones. As the result, the set of basic acoustical allophones appeared to be 3 times less than the number of articulatory allophones of English phonemes. The obtained theoretical results were used in solving the following applied problem. The assumption that certain groups of phonemes have similar coarticulatory effects on neighboring segments has made it possible to reduce the number of units in the inventory of the constructed sound database without affecting the quality of the synthesized speech. Time required for the formation of the inventory based on the set of basic acoustical allophones was reduced significantly. Besides the reduction of the number of basic units makes it possible to reduce time and effort in including new voices into the synthesis system. It should be also noted, that the obtained knowledge about the types of consonants having similar effects on neighboring sounds can be also successfully used in suballophone set formation. For example, in suballophone-based system the vowel halves with /w/ in left context can be used for constructing vowel allophones having both /w/ and a rounded vowel in left context. And, finally, the reduced set of basic acoustical allophones has allowed to develop compact speech segment database for English concatenative synthesis. Although the current trend in concatenative speech synthesis is to use large databases of acoustic units, smaller optimized databases are needed in low-memory devices (e.g. mobile phones, pocket PCs, etc.).

6. Conclusion

Thus the conducted evaluation tests demonstrated fairly high quality of the allophonic database inventory, which had been developed at the previous stages of our research. The perceptual tests showed that 85% of synthesized utterances were found by listeners practically indistinguishable from natural ones. At the level of segments the quality of the synthesized speech was evaluated as fairly high in terms of both naturalness and intelligibility. The junctions between allophones proved to be absolutely imperceptible. At present moment the inventory of the allophonic database for English concatenative speech synthesis contains 1200 elements (1000 vowel allophones and 200 consonant allophones). New voices can be built quickly with minimum human effort. The developed allophone database can be used in a complete English text-to-speech system.

7. References

5. O’Connor J.D. Phonetics. London (1977)