Allophone- and suballophone-based speech synthesis system for Russian

Pavel Skrelin

Department of Phonetics, St.Petersburg State University
Universitetskaya nab. 11, Saint-Petersburg, Russia
E-mail: paul@phonet.lang.pu.ru

Abstract. The speech synthesis system developed at the Department of Phonetics of St.Petersburg University combines the advantages of two methods of concatenative speech synthesis - diphone- and allophone-based ones. In a new synthesis we use physical realizations of allophones for the formation of the consonant inventory. The physical realizations of allophones are chosen with regard to their right phonetic context. In formation of the vowel set we used the method which combines allophone and diphone based synthesis principles: the database contains halves of vowel allophones (from the physical beginning of the allophone up to its middle and from the middle of the allophone up to its right physical boundary).

1 Introduction

The diphone-based system developed at the Department of Phonetics uses about 3000 units, the allophone-based system uses about 3300 physical realizations of allophones, but this quantity can be reduced by 30% without affecting the quality of the synthesized speech [3, 4]. The formation of sound units inventory requires considerable time and effort for speaker recording and material segmentation. This makes the introduction of new speakers' voices into the system very problematic, as this procedure is also laborious time-consuming. The reduction of the number of basic units makes it possible to reduce time and effort in including new voices into the synthesis system. The inventory of sound units in the newly developed speech synthesis system, which combines principles of allophone- and diphone- based systems, does not exceed 800; thus time required for the formation of the database is reduced 4 times.

2 Principles of Sound Unit Set Formation

A new speech synthesis system for Russian is based on allophone synthesis principles, e.g. on the using sound units, corresponding to the realizations of allophones in their natural physical limits. The difference of this system from the one developed earlier [3] consists in the following: in the process of the sound stream formation vowel allophones are constructed from vowel halves. The physical boundaries of vowel-halves (half-diphones or half-allophones) are chosen in such a way as to minimize spectral smoothing at the junction of formants and harmonics. Ideally the boundaries between vowel-halves are put in the middle of the vowel stationary part.

Thus, the inventory of sound units for speech synthesis includes:

1. Consonant allophones in their physical boundaries. For voiced consonants 3 allophones for each phoneme are required depending on the characteristics of the right context (before an unrounded vowel, before a rounded vowel, before a consonant). For voiceless consonants 5 allophones for each phoneme are required, depending on the characteristics of the right context (before an unrounded vowel, before a rounded vowel, before a voiceless consonant, a sonant and a pause). Sonants require 5 allophones each also depending on the properties of the right context (before an unrounded or a rounded vowel, before a voiced, voiceless consonant or a sonant, and before a pause). 11 allophones are required for /r/, considering its articulatory and acoustical properties and depending on its left and right phonetic contexts (between vowels, after a vowel before a voiceless consonant, after a vowel before a voiced consonant or a sonant, after a vowel before a pause, after a pause before a vowel, after a pause before a voiceless consonant, after a pause before a voiced consonant or a sonant, after a consonant before a vowel, after a consonant before a voiceless consonant, after a consonant before a voiced consonant or a sonant, after a consonant before a pause). It should be noted that for sonants a sound unit, corresponding to the realization of an allophone often performs a purely symbolic function, as its physical beginning may form part of the vowel allophone preceding it, while its end may be included in the vowel allophone following it [4].

2. Left vowel halves, which preserve physical characteristics determined by the left context. For the list of left contexts taken into account see [4].

3. Right vowel halves which preserve physical characteristics determined by the right context. Compared to the list of right contexts described in [4], the vowels before /r/ are added to the set.

At present the system uses 173 consonant allophones and 576 halves of vowel allophones. This number can be increased if we include fragments of theoretically possible, but exceptionally rare vowel allophones (for example, unstressed /o/) or consonants. Thus the whole number of sound units for synthesis will not exceed 800. It should be stated that the synthesis principle employed in the new system can be called allophonic only conventionally, as the structure of the sound units used is more complicated: some vowel allophones (and, consequently, their halves) include fragments of the preceding and following consonants; the beginning or end of the consonant allophone can be included in the vowel allophones; some of such consonants can be present in the database for the principle of the synthetic sound stream formation to be supported. For example, a post-stressed [j] is a 2 bytes signal of zero amplitude.

3 Vowel segmentation

The quality (naturalness) of sound of the constructed allophone depends on how accurately the boundaries between vowel halves are determined. In order to facilitate manual segmentation a procedure of mounting a signal on the bases of two fragments was included into EDS for Windows ver. 1.032 (a program for digital treatment of speech signals developed at St.-Petersburg State University of Telecommunication). This procedure allows the audio control of the mounting process and visual control of acoustic properties in the junction of two parts of the constructed signal in time and frequency domains. The segmentation process consists of three steps. The aim of the first step is to determine the physical beginning of second (right) vowel halves: for one (left) half-vowel taken from the most neutral left context (for example, after bilabials for rounded vowels or after velars for /a/), all second (right) halves are chosen (acoustic properties of right halves depend on the right context). The boundary place is chosen in such a way that physical positions of formants and harmonics were similar and corresponding to formant and harmonic values at the end of the used left half-vowel. Similarly, at the second step for one right half-vowel taken from the most neutral right context all left half-vowels are chosen. At the third step the remaining boundaries (between allophones) are determined. Segmentation made, the program allows listening to all variants of constructed vowel allophones. This procedure is used for determination of boundaries between vowels.

After segmentation the resulting vowel halves are normalized in amplitude and pitch is labeled. Pitch labeling is automatic, but the possibility of manual correction of inaccurately determined boundaries between Fo periods is provided. In the process of half-vowel set formation for speech synthesis we obtained an important theoretical and practical results. We found out that the position of a stationary part of the vowel [e] doesn't depend on softness of the left context. It means that the set of all necessary allophones can be formed with right vowel halves taken from vowels in hard left context only. These vowel halves can be used for constructing vowel allophones both in hard left context and in soft left context. A similar situation is found for the right vowel halves of [u], but their usage is restricted to unstressed vowels. It should be noted, that in diphone set formation for diphone synthesis system we have used two basic allophones for each vowel (in stressed and unstressed positions): one taken from the soft left context, the other taken from hard left context [1].

4 Modification of Physical Parameters of Allophones

4.1 Spectral smoothing in allophone forming

A possible lack of coordination in formants and harmonics positions at the junction of vowel halves in allophone formation requires the use of a special procedure of spectral parameters smoothing at the boundaries of vowel halves. The application of traditional spectral smoothing algorithms can result in distortion of individual characteristics of speaker's voice and in lowering of the naturalness of synthesized speech [4]. We tried to avoid it by maximally precise choice of vowel halves from the point of view of spectral parameters at their boundaries. As a result the effect of spectral smoothing is minimal and does not cause changes of individual characteristics of the speaker's voice. What is more, when the smoothing mode is off it does not lead to perceptible change of the synthetic voice quality. This fact demonstrates a high quality of the segmentation made. But as the possibility of difference in formants and harmonics positions at the half-vowels junctions exists, the necessity for smoothing remains. This procedure can become more required when using automatic means of sub-allophones extraction.

For spectral smoothing the procedure of audiomorphing [6] is used, but in calculations of initial and target spectra we use FFT instead of LPC which results in better sound quality. Some distortions in generated Fo periods are prevented by Hanning window treatment in time domain. This procedure is working in real-time mode.

4.2 Modification of Fundamental Frequency

In developing new speech synthesis system we decided not to use traditional Fo modification means in time domain. The new algorithm of Fo modification consists of three steps.

1. Pitch-synchronous spectra of initial signal are calculated. Then for each Fo period a signal presenting the sum of n Fo harmonics is generated (harmonics amplitudes is calculated on the initial spectrum basis). After that a new signal is pitch-synchronously subtracted from the initial one. This procedure results in residual signal without the initial Fo and harmonics.
2. A new signal is generated on the basis of calculated synthetic background: it presents the sum of n harmonics of target Fo. Amplitude values are calculated on the basis of spectra obtained in the first step.
3. Synthetic background is added to residual signal weighed in time domain.

To obtain a real-time mode operation of this algorithm, a part of calculations is implemented beforehand and calculation results are saved in a special database. Some data can be manually modified if necessary. Probably in the future by modification of this data values speaker's voice individual characteristics could be modified in the required direction. On the other hand, as the database contains the generalized data on formants and harmonics positions in allophones this information can be used for speech recognition tasks (knowledge-based, of course).

4.3 Duration Modification and Amplitude smoothing

The allophone duration is done in the same mode as in our previous system [3]. After all kinds of signal treatment, the amplitude smoothing at the boundaries of physical realizations of allophones is performed. To obtain it, the amplitude of the last n Fo periods of an allophone is gradually driven to the amplitude value of the first Fo period of the following allophone. As a basis for calculations of the energy of the first Fo period of the following allophone and n Fo periods energy values of preceding allophones are linearly interpolated to that value. Naturally if the allophone is followed by a pause or a voiceless plosive its amplitude is gradually driven to zero. This procedure is used only for vowels, voiced consonants and sonants, and it is working in real-time mode.

5 Text processing

Text processing means (stress marking, grapheme-to-phoneme conversion, intonation transcription etc.) were developed for our allophone-based speech synthesis system [5]. But in the last four years these means were improved to a great extent: the lexicon content was modernized and its volume optimized, the deepness of context analysis was increased, new pre-processing means were developed. For the time being the lexicon contains about 110 000 lexems giving more than 2 400 000 word-forms, about 800 000 of this quantity are proper names forms.

The lexicon structure was modified, too. For fulfill some tasks of the project "Segmentation of Continuous Speech as a Model of Language System Levels Interaction" (RFBR grant 98-06-80431) the transcription variants of some word-forms were introduced in the dictionary in addition to quasi-transcription of word-forms the pronunciation of which requires special rules. These transcription variants reflect more frequent deviations form the pronunciation standard. Such supplement allows modeling of non-standard but widespread pronunciation of frequent lexems. The use of the adaptive transcriber for grapheme-to-phoneme conversion [2] makes it possible to obtain the same result without adding special pronunciation variants to the dictionary. But the automatic transcription can not be limited to certain word-forms, as in this case all the word-forms in the text will obtain non-standard characteristics. On the other hand, the adaptive transcriber can easily provide different pronunciation variants to all lexemes in the dictionary; it allows the use of such dictionary for speech recognition tasks.

6 Conclusion

The described speech synthesis system for Russian showed good characteristics of intelligibility of synthesized speech and a higher level of its naturalness compared to the previous allophone-based system. Naturally this higher level of naturalness is due not to the new technology of allophone construction from two vowel halves, but to new methods of artificial modification of physical parameters of the sound stream. A much smaller number of sound units required for speech synthesis facilitates to a great extent the task of forming a basic set of sound units when new voices are added to the synthesis system. New methods of manipulation of spectral parameters of allophones now used for fundamental frequency modification can be used for obtaining new voices on the bases of one sound units set by artificial creation of the required characteristics in the spectra of sound units.

A new dictionary containing standard and different non-standard pronunciation variants of word-forms together with the database of formants and harmonics positions in the allophones set can be used in knowledge-based speech recognition systems.

References

1. Bondarko L.V., Kuznetsov V.I., Skrelin P.A. (1997) Zvukovaja sistema russkogo jazyka v svete zadach kompil'ativnogo sinteza. In: Bulleten' foneticheskogo fonda russkogo jazyka, N 6, St-Petersburg-Bochum, 60-84.
2. Shalonova K. (1997) Flexible Transcriber for Russian Continuous Speech. In: Proc. of SPECOM'97. Cluj-Napoca, 171-174.
3. Skrelin P. (1999) Allophone-Based Concatenative Speech Synthesis System for Russian. In: Proc. of the TSD'99 (Plzen), Berlin, 156-159.
4. Skrelin P. (1997) Concatenative Russian Speech Synthesis: Sound Database Formation Principles. In: Proc. of the SPECOM'97, Cluj-Napoka, 157-160.
5. Skrelin P. (1996) The Linguistic Software of the Concatenation Synthesis Systems for Russian Speech. In: Proc. of the SPECOM'96, St.-Petersburg, 165-166.
6. Slaney M., Covell M., Lassiter B. (1996) Automatic Audio Morphing. In: Proc. of the 1996 International Conference on Acoustics, Speech and Signal Processing, Atlanta.