АВТОМАТИЧЕСКОЕ СОЗДАНИЕ ТЕКСТОВЫХ КОРПУСОВ ДЛЯ ПОДГОТОВКИ ЗВУКОВЫХ БАЗ ГОЛОСОВ В СИСТЕМЕ СИНТЕЗА РЕЧИ НА РУССКОМ ЯЗЫКЕ

Соломенник А. И. (solomennik-a@speechpro.com),
ООО «Речевые технологии», Минск, Беларусь

Чистиков П. Г. (chistikov@speechpro.com),
ООО «Центр речевых технологий», Санкт-Петербург, Россия

Ключевые слова: синтез речи, unit selection, корпус текстов, дифон

AUTOMATIC GENERATION OF TEXT CORPORA FOR CREATING VOICE DATABASES IN A RUSSIAN TEXT-TO-SPEECH SYSTEM

Solomennik A. I. (solomennik-a@speechpro.com)
Speech Technology Ltd., Minsk, Belarus

Chistikov P. G. (chistikov@speechpro.com)
Speech Technology Center Ltd, St .Petersburg, Russia

This paper deals with the problem of speech database design for the needs of unit selection text-to-speech synthesis. An obligatory condition for the naturalness and intelligibility of synthesized speech is a high quality speech database. We propose a computer program developed specifically for the Russian language which creates a phonetically balanced text corpus of given size. We present a description of the program and a comparison of an automatically constructed corpus and some arbitrary corpora. The automatic text corpus generation program is part of a new voice building system for VitalVoice Russian TTS. It helps to supplement a text corpus with missing phonetic units. Further possible improvements of the algorithm are also discussed. We consider several ways to take into account intonational variation of units in a database at the stage of the preparation of a text corpus.

Key words: speech synthesis, unit selection, text corpus, diphone

1. Introduction

Unit selection synthesis is based on determining the best sequence of candidate units from the database. So the first step in developing a new synthetic voice is recording a speech database. Apart from the quality of linguistic processing or selection and modification algorithms, the naturalness and intelligibility of synthesized speech primarily depend on the quality of the speech database. For this reason, various possible units required to synthesize sentences of a natural language must be represented in the database.

There are several ways to deal with the situation when a required unit is missing in the database: simply skipping it; trying to replace it with one having similar characteristics (with possible modification); constructing it from smaller units, etc. But usually it causes a significant loss in naturalness in a particular place where an appropriate unit cannot be found.

The most common way to have all required units is to have large speech databases, representing dozens of hours of speech [1]. But only having a large amount of recorded speech is not enough, it should also be phonetically balanced, i. e. it should, if possible, contain all required units in all possible contexts with various possible characteristics such as acoustic parameters like fundamental frequency (pitch), duration, position in the syllable, and neighboring phones.

But since the creation of the database requires segmentation, which usually requires at least some manual correction after automatic segmentation ("forced alignment"), the size of the database influences the time necessary to prepare it for use. Besides, large databases are inconvenient to store and search in. So there should be a balance between the size and representativeness of a database.

Usually unit selection speech databases for high-quality TTS contain about 10 hours of recorded speech. Our experiments showed that 2–3 hours are enough for rather satisfactory quality and intelligibility. Less than one hour of recorded speech is insufficient because the necessary variability in such characteristics as pitch and duration will not be achieved, even if all possible units are present.

There are a number of investigations on the automatic corpora construction for different languages, including English, French, Chinese and other [2–5, 11–13]. But for the Russian language our work seems to be the first. The main advantage of our method is that it provides a convenient way for automatic voice creation by the possibility of not simple text selection from a large text corpus, but allows choosing a synthesis unit and creating and editing different preset corpora before selection.

2. Methods and Results

2.1. Instruments

Automatic text corpus generation software is part of the voice building system [8] for VitalVoice TTS [6] developed at Speech Technology Center. A new diphone

level was added to database segmentation [7], and at present the TTS engine selects a sequence of diphones for the target utterance. So the examples given in this paper refer to a text corpus for a diphone database, however our program allows obtaining statistics and creating corpora for allophone (i. e. triphone — allophone in context) units as well.

The automatic text corpus generation software was developed on the basis of a program performing the analysis of phonetic unit frequency [9].

Automatic transcription of texts is performed by a modified TTS engine with a disabled automatic break assignment module. Pauses are placed strictly at punctuation marks. This is important because the transcription depends on the positions of pauses. Therefore we imply that the speaker who is being recorded will make pauses in a similar way (it is necessary to control it during recording). It is more convenient than marking pauses in the text with special signs.

Our transcription system for Russian distinguishes 59 types of monophones: 19 vowels and 40 consonants (vowel position in relation to the stressed syllable is specified, and there are 4 additional voiced allophones for word boundaries). Consequently, the number of all possible diphones is 60*60–1 = 3599 (including the pause context). The number of all possible diphones permitted by the transcription rules (including nonsense words and combinations) is 2335. If various exceptions (such as “боа”) with unstressed /o/ and /e/ are taken into account the number increases to 2759. But not all these combinations are possible in natural human language. For example, only 2176 (from 2759 possible) combinations were found in a large text corpus containing over 460 000 word-forms [10].

2.2. Auxiliary corpora

In order to facilitate the text corpus design, the process is divided into three stages. They are built into the voice building system VoiceConstructor (Fig. 1), which includes all the stages up to installing the new voice.

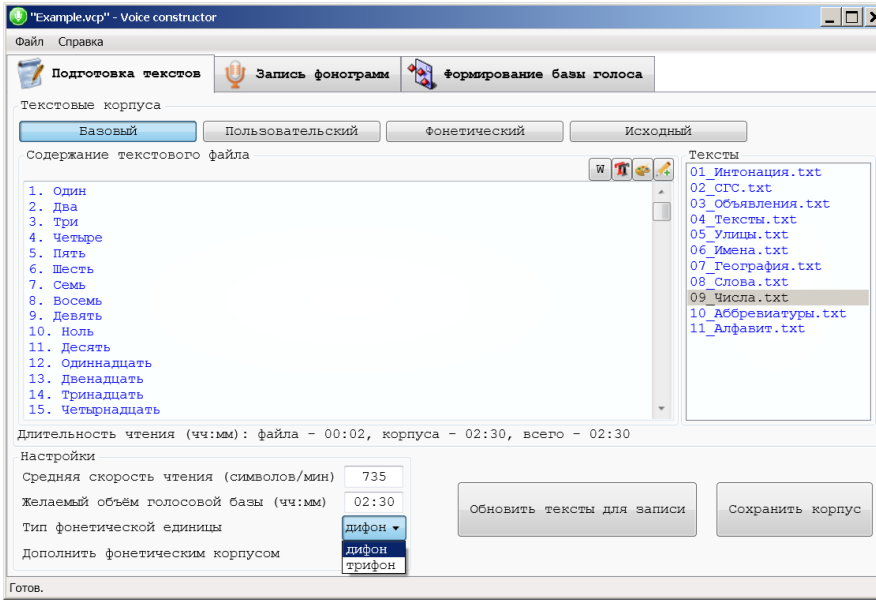

Fig. 1. The window of the basic text corpus in VoiceConstructor

Work with the system begins with specifying the settings of the future sound database. The user must choose the type of the basic unit: diphone or allophone, set the average speech rate (it can also be measured automatically during test recording) and the desired size of the sound database. The program shows the current size of the future database, durations of the current text and the current corpus. The current text is shown in the main window, a list of texts is on the right. The user can edit and delete each one of the texts.

There are four types of corpora: 3 (possible) parts of the resulting text corpus and one large text corpus for statistics and additional phrases:

1. Basic corpus
2. User corpus
3. Phonetic corpus
4. Large source corpus

Basic corpus

The basic corpus is the minimal corpus recommended for recording. It contains various frequent words and word combinations, texts with geographical and proper names, alphabet, abbreviations and some typical texts for speech synthesis applications (e. g. IVR or news). The total volume is 2.5 hours with an average speech rate. Any text can be disabled depending on the user's purpose.

User corpus

The user can upload any texts to the special user corpus. Usually these are phrases that the synthesizer is expected to be able to speak with high quality, e. g. welcome words or terms of use.

Phonetic corpus

The phonetic corpus is actually the main part of the resulting text corpus. It is the part generated automatically to include missing units in the database. It is constructed from sentences selected from the source corpus.

Source corpus

At first the source corpus was supposed to be a fixed large text corpus needed to obtain diphone statistics and to select phrases with missing units. But then we decided that it is more useful to allow users to change and upload texts. There are several reasons for it. First, for domain-specific TTS it is very important to choose sentences from a particular domain or genre. Apart from typical frequent words and combinations, we need to take into account possible changes in speech rate, manner and even pitch and timbre (compare reading books and announcements in an airport). Besides, a large corpus requires significant time for its processing. But the source corpus should be sufficient to represent the majority of possible units, and we recommend to include into it at least 10 hours of speech at an average speech rate (about 70 000 word-forms). An example of the source corpus which was compiled from different texts from web resources (news, politics, IT, fiction) is included to the system.

2.3. Algorithm

The algorithm for the generation of the phonetic corpus includes the steps shown in Fig. 2. First, the system transcribes all the necessary texts in a way described above. Then it calculates the desired volume of the phonetic corpus using the data about the total desired corpus size and the size of the basic and user corpora (if any).

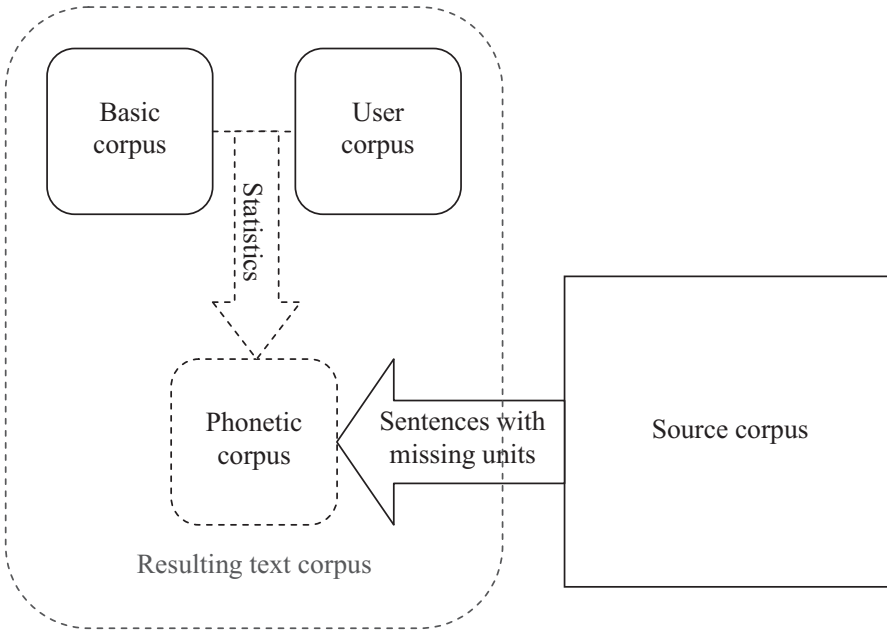

Fig. 2. Generation of text corpus

Sentences are chosen from the source corpus depending on how many missing units they contain, sentences with maximum missing units are taken first. If two sentences contain the same number of missing units, sentence with less frequent diphones will go first. The procedure is the same for infrequent diphones in the case when all diphones from the source corpus are already present in the basic and user corpora. The addition of sentences will stop when the resulting text reaches the desired size. This allows us to obtain a text of minimum size with maximum missing units. Of course the sentences in the text are not connected to each other in meaning, but each separate sentence is not artificial and easy to read. Speech databases usually consist of a number of files with only one sentence in each file, so sentences in VoiceConstructor are also recorded separately. Reading nonsense text also helps in controlling the emotions of the reader, as it is easier to read with constant rate and timbre.

After adding the phonetic corpus, all three corpora can be saved as text files and are displayed at the next stage when sound files are recorded.

2.4. Example of automatic text generation

To demonstrate the work of the program we compared a text constructed by the program to an arbitrary text or texts of the same size. We took relatively small amounts of data in order to speed up data processing. We generated a text for an hour of speech at the average speech rate (about 8000 word-forms) and compared its diphone and triphone statistics with texts of two different styles. The first was part

of the novel “Obmen” (The Exchange) by Soviet writer Y. Trifonov and the second was part of “Dialog 2010” proceedings (only Russian texts). Both also contained about 8000 word-forms.

Source text corpus contained more than 100 000 word-forms. It was drawn from different texts from web resources (politics, IT, fiction).

The statistics were obtained and compared by means of a specially designed program performing the analysis of phonetic unit frequency [9]. The results are shown in the Table 1 below:

Table 1. Diphone and triphone statistics in various texts

Text	Number of word-forms	Number of diphones (missing: in source/ in text)	Number of triphone (missing: in source/ in text)
“Obmen”	7835	1566 (5/462 — 23 %)	9852 (587/16054)
Dialogue 2010	8590	1488 (10/545 — 27 %)	8494 (415/17240)
Generated text	7557	1974 (0/49 — 2 %)	11 704 (0/13615)
Source corpus	106 260	2023	25 319

As we can see from the table, scientific text is statistically the poorest in phonetic variability, although it contains some unique diphones. The automatically generated text (we should note that it is generated in order to obtain diphone coverage, triphones were not taken into account) is the most representative and contains almost all the diphones presented in the source text. The table also demonstrates why diphone units are preferable to allophones: a context of a diphone in most cases is not as important as that of an allophone, so the required number of diphones is many times smaller.

A part of the resulting text corpus is presented below:

(1) *Еще более скептично смотрит на перспективу привлечения силовых структур к банковскому контролю эксперт «ФиэнЭкспертизы» Наталья Борзова: "Никакой потребности в каких-то дополнительных контролирующих структур сейчас нет, было бы куда полезнее, если бы ФСБ и МВД занимались своим делом, например, проверяли бы те адреса, по которым регистрируются фирмы-однодневки, тем более что силовые структуры и так уже имеют полномочия проводить выемку документов и получать всю необходимую для следствия информацию.*

Один из самых больших обманов 1 апреля, о котором долго потом вспоминали газеты и журналы, произошел в Лондоне в 1860 году, когда несколько сотен джентльменов с их чопорными английскими леди получили приглашение прибыть «на ежегодную торжественную церемонию умывания белых львов, которая состоится в Тауэре в 11 часов утра 1 апреля».

Антонио Вальдес-Гарсиа узнал из своей истории болезни, что ему были нанесены следующие телесные повреждения: закрытая черепно-мозговая травма, сотрясение головного мозга, закрытый перелом правой бедренной

кости с отрывом большого вертела, открытый двусторонний перелом нижней челюсти: центральный и суставного отростка слева со смещением, перелом альвеолярного отростка верхней челюсти слева, закрытый оскольчатый внутрисуставный перелом основания ногтевой фаланги 1 пальца левой стопы, множественные ушибленные раны лица, правого коленного сустава, левой стопы.

Президент Квасьневский, заметил Валенса, даже не смог достойно закончить свой президентский срок, помиловав экс-главу МВД и товарища по партии Збигнева Суботку, который был осужден на 3,5 года тюрьмы за передачу польской мафии секретной информации о готовящейся против нее операции (тем самым экс-министр поставил под угрозу жизни полицейских).

Он некоторое время слушал топот шагов убегающего Тигра после того, как тот скрылся за углом, как вдруг уловил звук дизельного двигателя средней мощности и сам припустил бегом по тротуару — как раз вовремя, чтобы увидеть задние габаритные огни грузовика, очень похожего на фургон, развозящий экспресс-почту, но оснащенного параболической антенной на крыше.

«Неформальных» разговоров в Сочи будет более чем достаточно, и, возможно, именно они будут играть ключевую роль, но, чтобы придать саммиту видимые результаты, стороны подпишут два соглашения в рамках одного из «общих пространств», как принято определять сферы практического российско-европейского взаимодействия: об упрощении визового режима для отдельных категорий граждан и о реадмиссии.

Что касается потенциальных «темных лошадок», то при подобном «уплотненном» сценарии, требующем уже на раннем этапе колоссальных затрат на рекламу и другую подготовительную работу в десятках штатов сразу, уверенно чувствовать себя может разве что нынешний нью-йоркский градоначальник мультимиллиардер Майкл Блумберг.

Здесь, правда, надо заметить, что и в прошлые годы страсти накалялись (в меньшей степени, конечно) именно к 16 декабря, поскольку в этот день 14 лет назад Генеральная Ассамблея ООН отменила свою же резолюцию 1975 года, объявлявшую сионизм формой расизма.

3. Conclusion

The program presented in this paper not only generates a phonetically balanced minimal text corpus but gives the user complete freedom in editing and choosing the type and style of the resulting text. It is very important when considering various possible applications of speech synthesis. But there are still some problems to solve.

One of the issues is when our sound database can be considered complete. A presence of only one realization of a particular diphone is usually not enough because it cannot cover all possible characteristics of a diphone such as pitch, energy

or duration without modification. For some units it is not as important as for others (for example, voiceless plosives vs. vowels under phrasal stress).

Natural sounding intonation is significant for high quality speech synthesis. It is very important to have units with at least two types of pitch movement to model phrasal and emphatic stress. The main tone movements normally appear on stressed syllables, so we may reduce the number of required diphones to those with stressed vowels (513 are possible). The intonation on them can be controlled by special signs for the speaker or simply by punctuation (full stop, comma or question mark). There are two possible solutions: automatically selecting sentences with the required punctuation and phonemic structure from a large text corpus, or manually constructing short sentences for the most frequent stressed units. The latter is more preferable since the former leads to a strong redundancy and some contexts (i. e. contexts with question marks) may not be present even in very large text corpora.

References

1. *Black A. W.* Perfect Synthesis for all of the people all of the time // Keynote, IEEE TTS Workshop. Santa Monica, CA, 2002.
2. *Bozkurt B., Ozturk O., Dutoit T.* Text design for TTS speech corpus building using a modified greedy selection. The 8th Eur. Conf. on Speech Communication and Technology (EUROSPEECH), Geneva, Switzerland, pp. 277–280.
3. *Chevelu J., Barbot N., Boeffard O., Delhay A.* Comparing set-covering strategies for optimal corpus design. Proceedings of the 6th International Language Resources and Evaluation, 2008.
4. *Chevelu, J., Barbot N., Boeffard O., Delhay A.* Lagrangian relaxation for optimal corpus design. Proceedings of the 6th ISCA Tutorial and Research Workshop on Speech Synthesis (SSW6). Bonn, Germany, 2007.
5. *Francois H., Boeffard O.*, Design of an optimal continuous speech database for text-to-speech synthesis considered as a set covering problem. Proc. Eurospeech, 2001
6. *Oparin I., Talanov A.* Outline of a New Hybrid Russian TTS System // Proc. of the 12th International conference on Speech and Computer, SPECOM 2007. Moscow, Russia, 2007. Pp. 603–608.
7. *Prodan A. I., Korolkov E. A., Oparin I. V., Talanov A. O.* MULTI-TIER MARKUP OF SPEECH CORPUS FOR HYBRID RUSSIAN TTS SYSTEM “VITALVOICE”. Komp’iuternaia Lingvistika i Intellektual’nye Tekhnologii: Trudy Mezhdunarodnoi Konferentsii “Dialog 2009” [Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialog 2009”]. Bekasovo, 2009.
8. *Prodan A. I., Talanov A. O., Chistikov P. G.* Voice building system for hybrid Russian TTS system “VitalVoice”. Komp’iuternaia Lingvistika i Intellektual’nye Tekhnologii: Trudy Mezhdunarodnoi Konferentsii “Dialog 2011” [Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialog 2010”]. Bekasovo, 2010.

9. *Smirnova N., Chistikov P.* Software for automated statistical analysis of phonetic units frequency in Russian texts and its application for speech technology tasks. Komp'iuternaia Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoi Konferentsii "Dialog 2011" [Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2011"]. Bekasovo, 2011, pp.632–643.
10. *Smirnova N., Chistikov P.* Statistics of Russian Monophones and Diphones. Proc. of the 14th International conference "Speech and Computer", SPECOM 2011, Kazan, 2011, pp. 218–223.
11. *van Santen, J. P. H., Buchsbaum A. L.*, Methods for optimal text selection. Proc. of Eurospeech, Rhodes, Greece, 1997, pp. 553–556.
12. *Wei Z., Yayu L., Ye D.* Automatic Construction for a TTS Corpus with Limited Text. Measuring Technology and Mechatronics Automation (ICMTMA), 2010.
13. *Wu H., Xu B., Huang T.* Automatic Corpus Selecting Algorithm Based on Triphone Models, Journal of Software, 2000, 11(2), pp. 271–276.