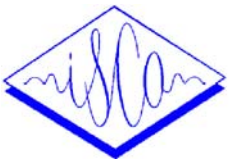
ISCA Archive
http://www.isca-speech.org/archive



EUROSPEECH '97
5th European Conference on Speech Communication and Technology
Rhodes, Greece, September 22-25, 1997

BELL LABORATORIES RUSSIAN TEXT-TO-SPEECH SYSTEM

Elena Pavlova, Yuri Pavlov, Richard Sproat, Chilin Shih, Jan P. H. van Santen

Bell Labs – Lucent Technologies
700 Mountain Avenue, Murray Hill, NJ 07974, USA
{yuriy,rws,cls,jphvs}@research.bell-labs.com

ABSTRACT

This paper describes the Bell Labs Russian text-to-speech system, a concatenative system with extensive text-analysis capabilities. The construction of Russian-specific modules will be discussed, including the text-analysis module, the acoustic inventory, the duration module, and the intonation module.

1. INTRODUCTION

The Russian Text-to-Speech system at Bell Laboratories represents one of the 11 languages/dialects in our multilingual text-to-speech system. All of these systems share the modular architecture given below, described in more detail in Olive and Sproat [4]:

- *Text Analysis:* Converts text into a linguistic analysis, including lexical properties of the words, and their phonetic transcription.
- *Duration:* Assigns duration values to each phone.
- *Intonation:* Generates fundamental frequency contours for sentences.
- *Amplitude:* Produces an amplitude contour for each sentence.
- *Glottal Source:* Generates glottal source parameters.
- *Unit Selection:* Converts phone strings into acoustic inventory elements, choosing the most appropriate ones from the inventory.
- *Unit Concatenation:* Generates LPC parameters and source-state information for the synthesizer.
- *Synthesis:* Reads the LPC parameters and source-state and outputs speech.

One important property of this system is that none of these modules are language specific. Rather they are table-driven, loading language-specific data at runtime. This property makes it easy to port the systems to different platforms and to adapt it for specific applications. Adding new languages requires constructing language-specific tables, but does not require one to write language-specific code.

2. TEXT ANALYSIS

The Russian TTS system uses the same architecture for text analysis as that of the German, French, Spanish, Italian, Romanian, Japanese and Mandarin systems. The model, which is described more fully in [6], uses *weighted finite state transducers* — WFST's — to map between the various levels of linguistic representation (orthographic input, lexical analysis, phonetic transcription) necessary for a linguistic description of the input. The WFST's are constructed using a lexical toolkit that allows declarative descriptions of lexicons, morphological rules, numeral-expansion rules, abbreviation expansion rules, phonological rules, and so on.

More specifically, the system works as follows:

- The input text is represented as a (trivial) unweighted acceptor I.
- Lexical analysis is performed using one or more lexical analysis WFST's L; $Lex = \pi_2[I \circ L]$ (the *right-hand projection* of the composition of I with L), defines a lattice of possible lexical analyses for the given input. Lexical analyses may have associated costs, meaning that some analyses will be cheaper (more favored) than others, in the absence of contextual information to decide among them.
- A set of 'language model' transducers Λ is composed with Lex to (partially) disambiguate lexically ambiguous forms using information from local context. The lowest-cost path of the resulting (partially) disambiguated lattice is then selected, to produce Lex', a unique lexical analysis.
- A set of one or more phonological transducers Φ is composed with Lex' in order to produce the phonetic transcription.

In the next few sections we present some details on each of these components. We first present the work on lexical analysis and phonological analysis, since these two components are the most complete to date. We then discuss prosodic phrasing, which is not currently incorporated into the system though it is clear how to do it. Lamentably, reasons of space do not allow us to discuss the complex and interesting problem of contextual homograph disambiguation.

10.21437/Eurospeech.1997-637

2.1. Lexical Analysis

An important distinguishing feature of our model of text analysis is that unlike most TTS systems, we do not distinguish 'text normalization' — e.g., expansion of abbreviations or digit strings into words — from the rest of linguistic analysis. So, the expansion of an expression like '20%' is handled as part of lexical analysis, just as is the analysis of an expression such as скидка (*skidka*) 'discount'.[1]

One reason that we eschew the more traditional model is that it simply does not work, and Russian provides some of the clearest evidence for this claim. In order to decide how to pronounce the abbreviation '%' in Russian, one generally needs to have a fairly detailed lexical analysis of the words in the surrounding context. For example in the expression *20%* скидка (*20% skidka*) 'twenty percent discount', we find the percentage expression modifying a following noun, in this case a feminine noun in the nominative singular. Following completely general grammatical principles of Russian, the percentage expression must in this case be in an *adjectival* form, and must agree in number, case and gender with the following noun: the correct form for '%' here is процентная (*procentnaja*) 'percent+adj+fem+nom+sg'. If, however, the percentage expression is not modifying a following noun, it must appear as a form of the *noun* процент (*procent*). Which precise form is used depends upon which number proceeds it, and the grammatical case assigned to the entire nominal expression. For instance, if the nominal expression '20%' occurs in a non-oblique case, then the word процент itself must occur in the genitive plural form процентов (*procentov*); if the expression were '22%' then the genitive *singular* form процента (*procenta*) would be required. Such a situation is hopeless for the traditional preprocessing model: instead one clearly must delay the decision on how exactly to 'normalize' a symbol like '%' until one has enough information to make the decision in an informed manner. Indeed, the lexical analysis WFST's for Russian simply transduce '%' into all possible renditions of that symbol, and it is the job of the contextual disambiguation ('language model') transducers to decide which is the correct form. This they do by modeling each of the cross-word grammatical cooccurrence restrictions: so, for instance, a nominal form of процент is filtered out when it precedes a noun.

As we have seen, selecting the correct form of an abbreviation in Russian depends on having a good lexical analysis of the context. More generally, one also needs morphological analysis in order to correctly pronounce Russian words. The main reason for this is that while stress is not normally marked in the orthography of Russian, stress placement is not 'regular' in that it cannot be predicted from the phonological structure of the word alone: rather it depends upon purely lexical features. For example, to know that the stress in the word карандашом (*karandashom*) 'with a pencil' falls on the last syllable, one needs to know that карандаш 'pencil' belongs to a class of nouns where the stress is placed on the case ending. Since unstressed vowels tend to be quite heavily reduced in Russian, misplacement of stress is quite noticeable; see Section . In order to avoid stress errors, one (minimally) needs a morphological analysis of all word — a non-trivial task in Russian.

Our morphological analyzer derives ultimately from the dictionary presented in [10]. Based on this dictionary, we produced three main databases. The first part is simply a list of stems tagged with relevant grammatical features, including information on inflectional class affiliation. The second is a set of tables — paradigms — expressing all of the endings possible for a given lexical class. The third database consists of morpholexical rules. The full description of Russian nouns requires 158 paradigms in our description; this large number stems from the combination of several critical factors, including the type of the stem (8), gender (3), animacy (2), and the accentuation type (10). Adjectives required 42 paradigms, and verbs 55 paradigms. Completely irregular forms were simply listed in separate files. All of this lexical information is compiled into a WFST that maps each ordinary word of text to its possible lexical analyses.

2.2. Word Pronunciation

The pronunciation of words in Russian depends upon two kinds of information. The first can be described as lexical information, and the second can be described in terms of regular phonological rules.

Lexical information affecting pronunciation includes lexical stress, as we have seen, but it also includes information on irregular or idiosyncratic spellings of certain sounds. The best-known case is the pronunciation of г (*g*) as /v/ when it occurs in the genitive endings его/ого (*ego*/*ogo*). Thus, земного (*zemnogo*) 'earthly+masc/neut+gen+sg' is pronounced as if it were written *zemnovo*. We code this kind of information *in the lexicon*, using special characters that for lack of a better name we term *archigraphemes*. The entry for the genitive ending in земного, for example, is given as о{g1}о, where the symbol {g1}, is mapped by the lexical transducers to orthographic г, but phonological /v/. To take another example, the consonant sequence вств (*vstv*) is pronounced sometimes as /stv/, sometimes (with regular devoicing of the first /v/) as /fstv/; the distinction appears to be lexically determined. Thus чувствовать (*chuvstvovat'*) 'to feel', is pronounced as if it were written *chustvovat'*, whereas in вдовствовать (*vdovstvovat'*) the *vstv* sequence is fully pronounced. We handle these 'silent' в's by representing them in the lexicon as {v1}, which maps to orthographic в, but phonological ∅. Similar 'archigraphemic' devices are used to describe other lexically conditioned pronunciations, including the many cases of foreign-derived words that have pronunciations that are not fully predictable from their spellings.

[1] See also [7], which independently proposes a similar model.

Regular phonological processes, including palatalization and voicing agreement in consonant clusters, reductions of unstressed vowels, and such miscellanea as the regular reduction of clusters like /stn/ to /sn/, are handled by rewrite rules, compiled into WFST's using the algorithm described in [2]. Information on both lexically conditioned and regular pronunciation was derived in large part from [1, 3].

2.3. Phrasing

The problem in prosodic phrasing prediction is to determine, for each word boundary, whether it represents merely a word boundary, or some higher prosodic break. Similarly, one needs in general to determine if a punctuation symbol (such as a comma) really corresponds to a phrase break or not. The problem is thus essentially one of lexical disambiguation, in this case of the material between words. Following [9], we have used a decision-tree based approach in investigating Russian prosodic phrasing. A 40 thousand word database was selected and hand-marked with the locations of weak and strong pauses. The factors used to train the tree include part-of-speech tags in a 4 word window, case tags in a 4 word window, accent information, punctuation, word length, sentence length, and the distance from the beginning and the end of the phrase. All length factors were coded in broad categories, otherwise there would be a vast number of gaps in factor combinations that could lead to poor discrimination on new data. In our data, we found that 1% of punctuation marks — specifically quotation marks, commas, and dashes — do not correspond to actual prosodic breaks; similarly 5% of prosodic breaks correspond to no punctuation mark. In cases where pausing is required but no punctuation mark is present, 84% of the cases happen when the sentence is at least 9 words long. The preferred location for phrase break is after a noun, the six highest ranking sites being between a genitive noun and a conjunct, a genitive noun and a past tense verb, a nominative noun and a preposition, a genitive noun and a preposition, a nominative noun and a conjunct, and an instrumental noun and a past tense verb. The prosodic phrasing prediction is not currently incorporated into the text analysis model. Doing so is nonetheless fairly straightforward, given the algorithm for compiling from decision trees into WFST's presented in [5].

3. ACOUSTIC INVENTORY

Like other Bell Laboratories TTS systems, the Russian synthesizer is a concatenative system using pre-recorded acoustic inventory elements. The current system distinguishes 54 phones, and has around 1700 acoustic inventory elements, which are primarily diphones.

We conducted several exploratory acoustic studies on the Russian speaker to determine the optimal inventory for the TTS system. Figure 1 shows the vowels and glides in the Russian sound inventory with formant information. The uppercase letters [A] [E] [I] [O] [U] represent soft (palatalizing) vowels, while the lowercase letters [a] [e]

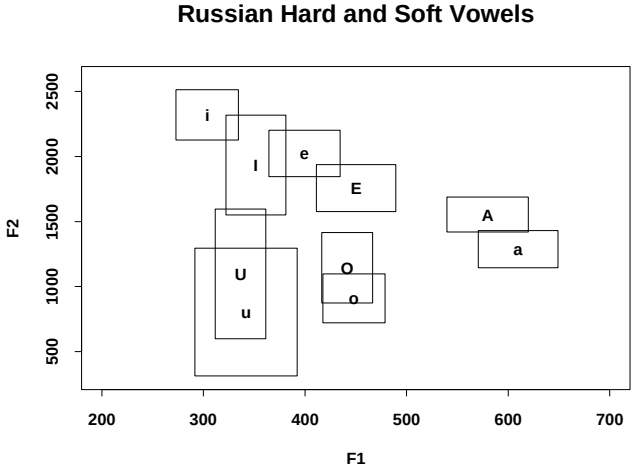


Figure 2: The soft and hard vowel contrast in Russian. Soft vowels are represented by uppercase letters while hard vowels are represented by lowercase letters. Vowel symbols mark median values (in Hz) in F1/F2 space. Boxes delimit standard deviations.

[i] [o] [u] represent hard (non-palatalizing) vowels.[2] [Y] is the onglide before soft vowels, and [y] is the offglide. Other symbols represent unstressed and reduced vowels. Unstressed orthographic <a> in the word final syllable is represented as [&]; unstressed orthographic <a> in other positions and unstressed orthographic <o> merge to [@]. Unstressed orthographic <e> in the word final position is represented by the caret, while unstressed orthographic <e> in other positions merges with unstressed [i].

Figure 2 shows a very consistent trend of centralization in the soft vowel series, in comparisons to corresponding hard vowels. Only stressed vowels are included in this plot. This tendency is most pronounced in [I], [E], and [A], where centralization is observed in both F1 and F2, and the vowel spaces of the soft and hard counterparts hardly overlap. For the rounded vowels [U] and [O], centralization is only observed in the F2 dimension. The consistency and the magnitude of the formant discrepancies in the Russian soft and hard vowels make it necessary to represent these two sets of vowels separately in the acoustic inventory.

Once the vowel and glide inventory is determined, target formant values are established for each of them via data analysis and listening tests. The inventory is chosen semi-automatically by ranking vowel candidates according to the goodness of their first three formants, i.e, how close they are to the target formants.

4. DURATION

Segmental durations are estimated from 31,000 phones recorded for the acoustic inventory database. Multiplica-

[2] In the TTS system the sound [i] is further split into stressed and unstressed versions, because the unstressed one is centralized. This point is not discussed in detail here for lack of space.

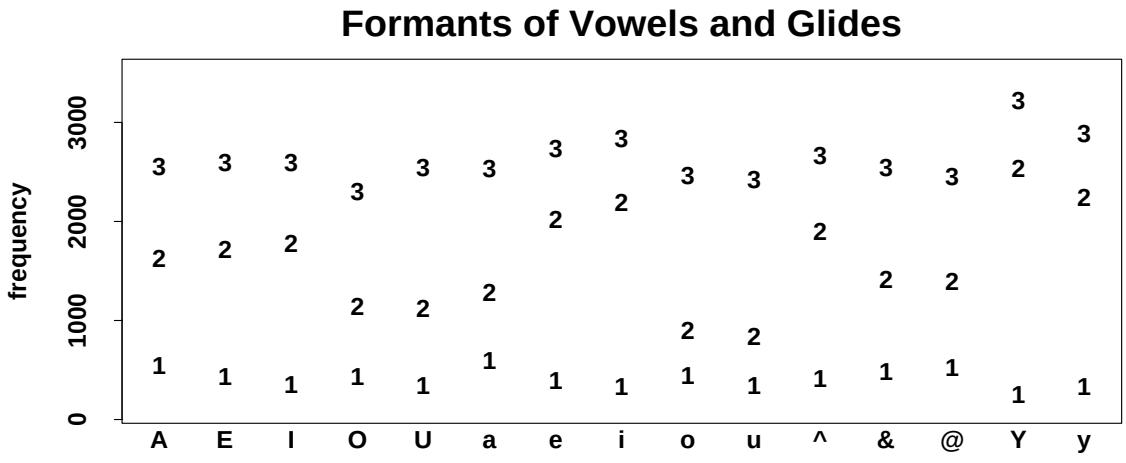


Figure 1: The vowel and glide inventory of the Russian TTS system. The sound symbols are listed on the X-axis, and numerals 1, 2, 3 above each symbol mark the median F1, F2, and F3 values for that sound.

i	u	o	e	a
119	119	130	131	144
I	U	O	E	A
113	127	130	126	138

Table 1: Vowel Duration in Msec

tive models are fitted separately for major classes of sounds, such as vowels, the closure portion of stops and affricates, the burst portion of stops and affricates, fricatives, sonorants, and trills. The independent factors used to estimate segmental duration includes the identity of the phone, the major phone class of the previous and the following phone, syllable types, stress, and positional factors including the distance of the phone from the stressed syllable, and whether the phone in question is in the initial or final position of a syllable, a word, or an utterance. The overall correlation between the observed and the predicted duration is 0.73, and the rms is 22 msec.

Table 1 gives some of the estimated vowel duration in msec, which are corrected for the effect of factors. The duration correlates inversely with vowel height: low vowels tend to be longer, while high vowels tend to be shorter. This is a tendency observed in many languages, including English.

5. INTONATION

The intonation module assumes the model described in [8], where the intonation contour is computed by adding accent curves to the phrase curve. The accent curve is anchored on an accent group, a unit consisting of a stressed syllable and all following unstressed syllables. Both the peak location and the shape of the curve are estimated from weighted durational parameters: the duration of the stress group, the duration of the onset, and the duration of the sonorant portion of the rhyme. The shape of the accent curve can be warped by changing the weights, and the pitch height can be shifted up or down. In Russian, nonfinal accent curves have a more gradual falling slope than nucleus accent curves. Also, in comparison to English, the peak placement in Russian accent curve is earlier.

6. REFERENCES

1. R. I. Avanesov (ed.) Орфоэпический словарь русского языка (*Orthoepic Dictionary of the Russian Language*). Moscow, Russkij Yazyk, 1983.
2. Mehryar Mohri and Richard Sproat. An efficient compiler for weighted rewrite rules. In *34th Annual Meeting of the Association for Computational Linguistics*, pages 231–238, Morristown, NJ, 1996. Association for Computational Linguistics.
3. R. F. Kasatkina and M. L. Kalenchyk. Материалы к орфоэпическому словарю (*Materials for an Orthoepic Dictionary*), Moscow, 1995.
4. Joseph Olive and Richard Sproat. Text to speech synthesis. *AT&T Technical Journal*, 74(2):35–44, 1995.
5. Richard Sproat and Michael Riley. Compilation of Weighted Finite-State from Decision Trees. In *34th Annual Meeting of the Association for Computational Linguistics*, pages 215–222, Morristown, NJ, 1996. Association for Computational Linguistics.
6. Richard Sproat. Multilingual text analysis for text-to-speech synthesis. *Journal of Natural Language Engineering*, 1997. to appear.
7. Christof Traber. SVOX: The implementation of a text-to-speech system for German. Technical Report 7, Swiss Federal Institute of Technology, Zurich, 1995.
8. Jan van Santen. Segmental duration and speech timing. In Y. Sagisaka, N Campbell, and Higuchi N., editors, *Computing Prosody: Computational Models for Processing Spontaneous Speech*. Springer, New York, 1997.
9. Michelle Wang and Julia Hirschberg. Automatic classification of intonational phrase boundaries. *Computer Speech and Language*, 6:175–196, 1992.
10. A. A. Zaliznjak. Грамматический словарь русского языка (*Grammatical Dictionary of the Russian Language*). Moscow, Russkij Yazyk, 1977, (Machine-readable version, 1991).