PHONETICALLY SUFFICIENT ALLOPHONIC DATABASE FOR CONCATENATION SYNTHESIS OF RUSSIAN SPEECH

Nina V. Zinovieva
Moscow State University, Moscow, Russia

ABSTRACT

The paper describes a phonetically sufficient database of Russian speech samples corresponding to phoneme-size units (or so called allophones) derived from different phonetic contexts. The database was designed for the purpose of automatic speech synthesis and was implemented in a concatenation text-to-speech Russian synthesis system

INTRODUCTION

The goal of the research is to create an optimal database (or inventory) of speech units for a Russian text-to-speech synthesizer, based on waveform concatenation. During the work we had to solve several problems. One of them is the appropriate choice of basic concatenation units. They may be diphones, triphones, syllables, demisyllables, and even words. However, all of them usually modify their acoustic quality in speech string (due to the coarticulation influence of adjacent elements). It seemed reasonable to choose the allophone (which we understand as *acoustically and perceptually distinguished context dependent realizations of phoneme)* as the basic unit, so that we could model the coarticulation phenomena when these units are being spliced together to form the synthesized speech.

The choice of linguistically motivated units (allophones) enables (a) to cluster phonemes in classes according to their flexibility in continuous speech, and (b) to cluster different context environments in groups according to their influential power and type of influence. Using these two features of allophonic approach we created an optimal set of speech units (a total of 667, using about 1 Mb memory to store them) which covers all coarticulation effects and thus provides the natural sounding of the synthesized speech.

The basic method for elaborating an exhaustive inventory of Russian allophones consisted of (a) an expert estimation of extracted sound wave segments corresponding to allophones and (b) an expert estimation of allophones extracted from one context and implanted into another, with similar or different coarticulation effect.

As a rule, the units for the concatenation are phoneme-size segments of the speech wave, although there are some exceptions. For example, to create stops, affricates and trills [r] and [r'], usually more than one acoustic segment are used; while to synthesize some two-phoneme sequences, such as post-stressed endings, a single acoustic segment can be used.

The main result of the work is a full description of the Russian allophonic system and elaboration of the optimal allophonic inventory in the form of acoustic speech signal (wave-form).

PHONETIC INVENTORY

The necessary condition for the work of an allophonic processor is the availability of phonemic transcription of the text to be synthesized. It is provided by an automatic transcriber specially designed for the Russian speech synthesis system. We used the following inventory of the Russian phonemes.

1. Stressed vowels: [á], [ó], [ú], [é], [í], [ý];
2. Unstressed vowels of the first degree of reduction: [a], [u], [i], [y] [o], [e]. The last two unstressed vowels are not regularly used in standard Russian, but sometimes they are pronounced in borrowed words.
3. Unstressed vowels of the second degree of reduction: [ax], [ix], [ux]
4. Non-palatalized consonants: [p], [t], [k], [b], [d], [g], [s], [sh], [z], [zh], [f], [v], [x], [c], [dz], [γ], [m], [n], [r].
5. Palatalized consonants: [p'], [t'], [k'], [b'], [d'], [g'], [s'], [z'], [sh'], [zh'], [f'], [v'], [x'], [ch'], [dzh'], [m'], [n'], [r'], [j']

One can see that the phonemic inventory used in our work slightly differs from that prevalent in Russian phonetic descriptions. This is because, for the purpose of synthesis, we had to choose such units that not only represent the phonemic relationships but also have acoustic and perceptual identity. It means that we have different units in our phonemic transcription even for those pairs that are in no meaningful contrast, but nevertheless, have different acoustic patterns which cannot be derived from one another: [x-γ], [c-dz], [ch'-dzh'], unstressed vs. stressed vowels, etc.

BASIC CONCATENATION UNITS

In arranging our database we proceed from the following three assumptions:

1. the amount of context-dependent variants is significantly larger for vowels than for consonants;
2. different consonants are affected by context influence to different degrees;
3. because of the prevalent CV-type of the Russian syllable, the left context is more important for vowels while the right context is more important for consonants.

According to these assumptions and a vast amount of preliminary expert estimations of the phoneme-size wave segments taken from different contextual environments, we divided the set of phonemes into the following classes.

Classes of vowels

1. Stressed vowels: [á], [ó], [ú], [é], [i], [ý];
2. Unstressed vowels: [a], [u], [i], [y], [o], [e], [ax], [ix], [ux];

Classes of consonants

1. Non-palatalized stops: [p], [t], [k], [b], [d], [g];
2. Palatalized stops: [p'], [t'], [k'], [b'], [d'], [g'];
3. Non-palatalized non-velar fricatives and affricates: [s], [sh], [z], [zh], [f], [c], [dz];
4. Palatalized non-velar fricatives and affricates: [s'], [z'], [sh'], [zh'], [f'], [v'], [ch'], [dzh'];
5. Nasals: [m], [n], [m'], [n'];
6. Liquids and velar fricatives: [l], [l'], [v], [v'], [x], [γ], [x'];
7. Trills: [r], [r']
8. Glide [j']

For each class the following relevant contextual environments were determined which affect the phonemes of the class, creating their allophonic modifications.

Relevant contexts for vowels

A. Left contexts for vowels

1. Beginning of syntagm-initial word.
2. Dental and alveolar non-nasal non-palatalized consonants, and central vowels: [d], [t], [s], [z], [c], [dz], [sh], [zh], and [á], [é], [a], [e], [ax].
3. Labial non-nasal non-palatalized consonants, and labialized vowels: [b], [p], [f], [v], [m], and [ú], [ó], [u], [o], [ux].
4. Velar non-palatalized consonants: [k], [g], [x], [γ].
5. Dental nasal non-palatalized consonant: [n].
5. Labial nasal non-palatalized consonant: [m].
7. Non-palatalized trill [r].

8. Non-nasal palatalized consonants, and front vowels: all consonants marked with the palatalization symbol (except [n'] and [m']), and vowels [i], [ý], [i], [y], [ix].

9. Dental nasal palatalized consonant [n'].

10. Labial nasal palatalized consonant [m'].

B. Right contexts for vowels

According to the third assumption, only five types of right contexts were considered for vowels:

1. End of syntagm-final word.

2. Non-labial non-palatalized consonants, and central vowels: [d], [t], [s], [z], [c], [n], [dz], [sh], [zh], [k], [g], [x], [γ] (the last four, only when not followed by labialized vowels [ú], [ó], [u], [o], [ux]), and [á], [é], [a], [e], [ax], [ý], [y].

3. Labial non-palatalized consonants, labialized vowels, and velar consonants followed by labialized vowels: [b], [p], [f], [v], [m], [ú], [ó], [u], [o], [ux], and [k], [g], [x], [γ] followed by [ú], [ó], [u], [o], [ux].

4. Non-palatalized thrill [r].

5. All palatalized consonants, and front vowels [í], [i], [ix].

Relevant contexts for consonants

For different groups of consonants different sets of relevant contexts were determined, to minimize the allophone inventory and, respectively, the database. This was achieved due to the fact that different consonants are differently sensitive to the environment, so they have different numbers of allophones. This difference will be shown in Table 1.

In compliance with the above, let us divide the relevant contexts for consonants into the following groups

A. Groups of left contexts for consonants

1. Beginning of syntagm-initial word.

2. Labialized non-reduced vowels [ú], [ó], [u].

3. Labial consonants, and labialized non-reduced vowels [b], [p], [f], [v], [m], and [ú], [ó], [u].

4. Front vowels: [í], [ý], [i], [y], [ix].

5. Front vowels [í], [ý], [i], [y], [ix] and all palatalized consonants.

6. Other left contexts

B. Groups of right contexts for consonants

1. End of syntagm-final word.

2. Labialized non-reduced vowels [ú], [ó], [u].

3. Labial consonants, and labialized non-reduced vowels: [b], [p], [f], [v], [m], and [ú], [ó], [u].

4. Any non-final context for palatalized consonants.

5. All vowels.

6. Other right contexts

Table 1 shows what groups of left and right contexts are relevant for each class of consonants.

The table does not show that there are special intervocalic allophones for consonants of the seventh group (trills [r] and [r']). Each of these intervocalic allophones consists of a single element. When the trills occur not intervocalically, the corresponding context-dependent allophone is added to the intervocalic element from the side of the consonant neighbor. If both neighbors are consonants, trills are synthesized from two context-dependent segments and intervocalic allophone between them.

DATABASE PREPARATION

To prepare the allophonic database, a special vocabulary was compiled of words containing all the necessary allophones in positions convenient for cutting off. When selecting the environments, only one typical representative of each group of contexts

Table 1. Left and right contexts relevant to different groups of consonants

Classes of consonants	Groups of left contexts						Groups of right contexts					
	1	2	3	4	5	6	1	2	3	4	5	6
1						+	+	+				+
2						+	+	+		+		+
3	+					+	+	+				+
4	+					+	+			+		+
5	+					+	+					+
6	+	+		+		+	+	+		+		+
7	+		+		+	+	+		+	+		+
8	+	+			+	+	+				+	+

was taken. The selected words were pronounced and recorded, one by one, by a professional TV-announcer. Then, the recordings were digitized and, with the help of a special sound processor, the allophones were cut off and stored in the database as separate files. Each file was given a name containing information about the group to which the phoneme belongs, its individual identity, and both right and left contexts from which it was taken and to which it should be implanted during the synthesis procedure.

SYNTHESIS PROCEDURE

During the synthesis, the input text is automatically transcribed, then the chain of transcription symbols is converted into a sequence of allophones corresponding to the context in which they occur The allophone are labeled by the above mentioned file names, and by these names they are extracted from the database and spliced together. Specia rules for pitch and duration patterns are implemented to make the synthesized speech sound as natural as possible.

ACKNOWLEDGMENTS

The author thanks I.Frolova and L.Zacharov for their assistance in preparing database, and Dr. Vladimi Segal for his helpful suggestions.