A Study of the Acoustic Model Choice for Russian Speech Recognition[1].

Chuchupal Vladimir, Makovkin Konstantin, Gorokhovsky Konstantin, Chichagov Alexander

Computer Center, Russian Academy of Science, Vavilova 40, Moscow 117967, RUSSIA
chuchu@ccas.ru, makovkin@ccas.ru, kgor@infratel.com, tchichag@ccas.ru

Abstract: The results of the ongoing work on acoustic modeling of Russian are presented. The basic approach for the modeling is a decision tree technique applied to the conventional context-dependent HMM phone models. We reworked and extended substantially our questionnaire (more than 100 new questions) to include more detailed and, possibly, important phonetic events. The global allophone trees for Russian language have been built and discussed. The structure of two gender-dependent allophone trees is compared. The list of most important questions is defined and investigated from the view of the existing phonetic knowledge.

1. INTRODUCTION

We present the results of our ongoing efforts on acoustic modeling of Russian spoken language. The primary target was to elaborate the compact and accurate enough set of model for telephone connected speech recognizer with small vocabulary of digits and control words. The basic approach for elaboration of the phone set was a binary decision tree clustering algorithm [1].

The decision tree construction includes the choice of the:

- The set of the root nodes (or just one node for the global tree);
- The set of questions, concerning the phone identity, state identity, left or right context identity, etc.;
- A node splitting criteria, that establishes the rules accordingly to these the question (from a question set) and the node ("parent" node from the tree nodes) are chosen to split the parent node into the couple of the new child nodes. The node splitting criteria used here was based on the log likelihood estimate for the observations belonging to the node.

The initial set of nodes (the roots of the decision tree) in the experiments was the conventional phones for the multiple node trees and a single "catch-all" node for the global decision tree.

There are two questionnaires have been used. The first one have been used for the global trees experiments and the second questionnaire have been used for the trees with multiple roots. The global tree questionnaire was different in that it contains a number of questions for identification of the central phone. Table 1shows some questions (that used below in the paper) .This table contains the description of questions (only for the left contexts, the right context and center questions are the similar) and the question's short form as it was used in the other pictures and tables below. Prefix L_ denotes questions to the left context and prefix R_ denotes the questions to the right context:

Is left context a shifted?	L_Soft ?
Is left context a sonorant?	L_Snr ?
Is left context a lips consonant?	L_ConLips?
Is left context a two-focus?	L_2Focus?
Is left contest a forward vowel?	L_Forw?
Is left contest a forward sonorant?	L_Forv?
Is left context a labial?	L_Labial?
Is left context a low vowel?	L_VLow?
Is left context a "sz" sonorant?	L_Svist?
Is left context a mid vowel?	L_VMid?
Is left context a round vowel?	L_Round?
Is left context a high vowel?	L_VHigh?
Is left context a back consonant?	L_Back?
Is left context a round vowel?	L_Round?
Is left context a vowel?	L_Vow ?
Is left context the "l" phone?	L_Plos?
Is vowel a stressed?	STRESS?
Is sonorant a shifted?	SHIFT?
Is vowel if under reduction?	REDUC0?

The database used for the design of phone inventory was the bootstrap part of the Russian acoustic-phonetic database for the telephone applications [1].

The system front-end processor converts the input signal into four vectors of features: spectrum, delta-spectrum, energy and delta energy, calculated in 16 frequency bands (equally spaced in the Mel scale). For discrete density HMM based decision tree the incoming short time parameters are then coded with codebooks that in turn represent the 2D self-organized feature map of equal 29*29 dimensions.

Modeling of speech signal for the decision tree relies on the discrete left to right HMMs for phone representation, with 3 states per phone with null language model.

The Section 2 describes the decision tree construction procedure in terms of the question set, the initial set of the root nodes and node splitting criteria.

[1] The work was supported in part by Russian Basic Research Fund, grant # 02-01-00453

The Section 2 describes the experiment with global decision tree. The Section 3 describes the gender dependent decision tree.

The Section 4 describes the recognition accuracy for the small vocabulary task and different sizes of phone set.

2. GLOBAL DECISION TREE

The purpose of the global decision tree was to understand if usual implementation of tree growing procedure with multiple roots that are chosen in the conventional phones is worth from the point of global tree, or we should choose the initial roots in some other manner.

The table below describes the 30 first applied questions (independently of the model state) together with their relative importance (in terms of scaled score gains, relevant to log likelihood). The question showed in the leftmost column, and the total accumulated gain estimation is showed in the right most column

Question	Meaning	Gain
Vback	Is center a back vowel?	16990
VceLes	Is center a voiceless?	11356
Nasal	Is center a nasal?	5130
Plos	Is center a plosive?	5110
Forv	Is center a forward sonorant?	4894
Vlow	Is center a low vowel?	4433
Snr	Is center a sonorant?	3745
Vforw	Is center a forward vowel?	3103
SHIFT	Is center a shifted?	3015
2Focus	Is center a two-focus?	2758
Svist	Is center a “SZ” phone?	1910
Lips	Is center a lips?	1612
Back	Is center a back consonant?	1814
R_Soft	Is right context a shifted?	1618
SHIFT	Is center a shifted?	1602
D	Is center a D phone?	1513
Svist	Is center a “SZ” phone?	1297
CH	Is center a CH phone?	1249
L	Is center a L phone?	1067
TS	Is center a TS phone?	859
Back	Is center a back sonorant?	880
R_Soft	Is right context a shifted?	861
F	Is center a F phone	772
Vmid	Is center a middle vowel?	723
X	Is center a X phone?	677
STRESS	Is center a stressed?	646
L_VceLes	Is left context a voiceless?	571
SHIFT	Is center a shift?	548
SHIFT	Is center a shift?	544
L_Soft	Is left context a shifted?	540

Table 2. First 30 questions for the global decision tree.

Accordingly to expectations, the questions about quality of the central phone are dominant in the beginning of the tree growing, while the questions for the left or right context become important later. Among first 30 questions (see table above) there are 4 the questions concerning the left or right contexts and 26 questions about the central phone, for the second 30 questions there are 12 questions concerning a context and for the next 30 questions there already 23 questions concerning the left or right contexts, while only 7 questions concerning the central phone.

Is was expected that some sounds (like voiceless plosive P, T, K; nasal N, M; voiced B, D, B; etc.) possibly would not be described in terms of different leaves of the decision tree. However it appears that all of these phones have been associated with the different tree leaves very soon after the tree grow procedure begun.

Central hard M and N have different tree nodes for the right context of back vowel. That is both those sounds in central position are described by the same (central) allophone with the derivative name of NotVBackNotVceLesNasalNotSHIFT. There are two different allophones for central M and N if the right context is a back row vowel. In such a case the center m and n are distinguished with the decision tree so the different phone models existed. The situation is illustrated with the following Figure 1.

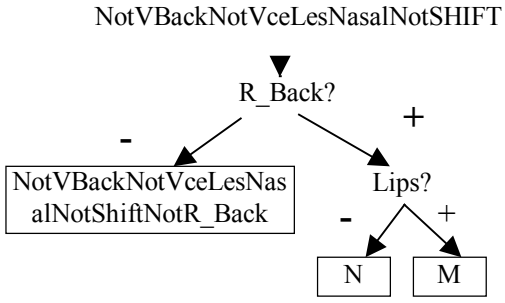

Figure 1. Allophones for central M and N.

The hard phones (as central phones for the triphone) P, T and K were successfully separated during tree grow and are represented with different leaves as it describes below:

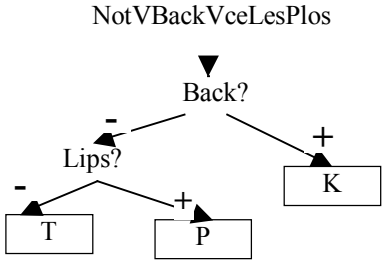

Figure 2. Allophones for central P, T and K.

The hard phones B, D and G also have been represented with the different leaves as it describes below:

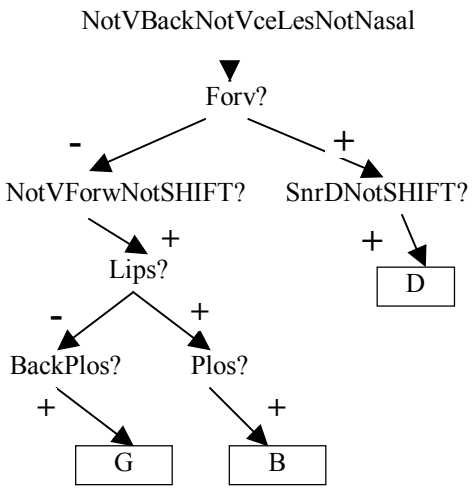

Figure 3. Allophones for central B, D and G.

3. Gender Dependent Decision Trees

The gender dependent tree has been grown with multiple initial roots. The standard phone set has been used to serve as initial tree root node set. It is known that the gender dependent models provide us with the higher speech recognition rate. The issue was, could we use the same allpohone set for male and female models, or it worth to use the different alphabets depending on the dictor's gender. Table 3 and 4 below describes the most important questions for male and female trees. The question rank means relative frequency of question use (during approximately first 120 most important questions). The questions following in the order of their rank, not the gain score. Also the questions with rank 1 or less are omitted.

Male Tree		
Question	Rank	Gain Score
SHIFT	14	3654
R_Soft	7	1521
STRESS	7	862
L_Low	6	470
R_Forv	6	377
L_High	5	319
L_Nasal	4	475
L_Soft	4	346
R_Round	3	350
R_Labial	3	278
L_Labial	3	265
L_Svist	2	761
R_ConLips	2	254
R_VoiceLess	2	212
R_L	2	213
L_ConLips	2	188
L_L	2	185
REDUC1	2	161
R_Low	2	144
R_Sonant	2	138

Female Tree		
Question	Rank	Gain Score
SHIFT	15	4088
R_Soft	7	1788
STRESS	6	1069
L_Low	6	479
L_Soft	5	667
L_VoiceLess	4	566
L_Nasal	4	295
R_Sonant	4	258
R_Low	4	242
R_Forv	4	88
L_ConLips	3	357
L_High	3	233
R_VoiceLess	3	229
L_Labial	3	185
REDUC1	2	431
R_Round	2	290
R_Labial	2	178
R_L	2	159
REDUC0	2	145
L_Sonant	2	114

Table 3. First high-ranking questions for the gender dependent decision tree.

As it seen from the Table 3 the resulted triphone alphabets for male and female are not strictly identical, however, the difference is not large. There is one question (namely "is left context a voiceless?") for the female tree that is high ranking for the female and not high ranking for the male tree. However, in large part the both trees are similar and the existing difference could be explained by the speech material features.

It is worth to note that behavior of gain score value during tree grow procedure was not strictly monotonic decrease function like it was expected. The graph below depicts score gain value on the first 100 cycles for the global tree. We explain the non-monotonic behavior by the hidden regularity in parameter's distribution (for example, see Figure 4 at step 11).

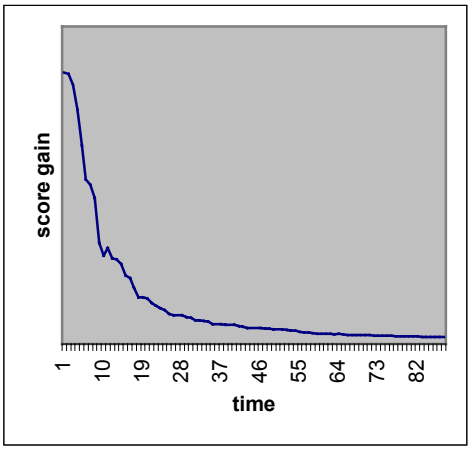

Figure 4. The graph of score gain value for the tree grows procedure

4. WORD RECOGNITION EXPERIMENTS

The following table contains the comparative estimates of recognition accuracy (for discrete word recognition, null grammar for selected words, the vocabulary size was about 50 words) depending on the phone alphabet size. This is not the best word recognition results for this vocabulary and our system but the comparative results for different phone set sizes. There are outlined the results when the recognized words have been comprised of the "known" phone models (i.e. that actually have been met during model training) and results when the recognized words have contained the "invisible" phone models (the type that never have been met during training). Also the results when the speech samples of the dictor has been used for model adaptation are presented ("known" dictor).

Phone set size	Word error rate			
	Known dictor and models	Unknown dictor and known models	Known dictor and unknown models	Unknown dictor and models.
100	1%	6%	11%	20%
200	1%	6%	10%	16%
400	4%	4%	11%	20%
600	4%	4%	11%	20%

Table 4. Word recognition error rate for different alphabet sets.

The maximum robust (that is stable enough not to be removed during pruning procedure) number of allophones for the speech corpus used was about 700 (approximately 2K HMM state models). However the optimal number of phone models for test vocabulary recognition appears to be about 200 phones. We explain such an amount of models from the point of size our speech database and the size of test vocabulary.

CONCLUSION

The results of the ongoing work on acoustic modeling of Russian are presented. We reworked our questionnaire and consider gender dependent phonetic decision tree and global phonetic decision tree for Russian. The results of word recognition experiments for small vocabulary task are presented.

LITERATURE

1. L.Breiman, J.Friedman etc. "Classification and Regression Trees", Wadsworth, Inc, 1984.

2. S.Gelfand, C.Ravishankar, and E.Dell, "An Iterative Growing and Pruning Algorithm for Classification Tree Design" IEEE Trans, PAM V13, nol2, pp.163-174, Feb.1991.

3. Design and implementation of the Russian telephone speech database. In Proc. of Int. Workshop "Speech and Computer", SPECOM'99, Moscow, 1999, pp.179-181.