[1]Accurate acoustic modelling for Russian.

Chuchupal Vladimir[1], Makovkin Konstantin[1],Chichagov Alexander[1]
chuchu@ccas.ru, makovkin@ccas.ru, *chichag@ccas.ru*
Computer Center, Russian Academy of Science, Vavilova 40, Moscow 117967, RUSSIA

Abstract: The results of experiments on acoustic modeling for Russian are presented. The primary purpose was to build the accurate and manageable set of phone models for the public telephone network continuous speech recognizer. The set of binary questions relating to the context, phone and state identity was elaborated and implemented via a decision tree technique to the different initial sets of phones. The particular phone inventory depends on the node splitting criteria, type of speech material (telephone speech or microphone quality speech), and type of HMM models used (discrete vs. semi-continuous).

1. INTRODUCTION

To develop a connected speech recognition system on the base of context-dependent phone units we need a manageable (relatively small size) set of context – dependent units. From the point of operation speed and memory requirements the desired number of models for our case were limited to 1000 units.

There are two very popular approaches to elaborate the set of phone models for speech recognizer. The first approach is an automatic manner way to build the phone models by automatic cluster procedure with the succeeding splitting the models into the smaller ones until the termination condition has been reached. The other well-known approach is the decision trees. In our case the only advantage in implementing the decision trees approach was that we expect to utilize its generalization capabilities. Decision tree also provide the designer with possibility to model (via tying mechanism) so called "unseen" phones. This was important in our case since it appears that the bootstrapping part of the database (that was used for initial learning) did not contain some phones that have been required by some vocabularies.

The HMM type for the test below were the discrete ones, with 2D codebooks build to be Kohonen's self-organized feature maps

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

The Section 3 describes the phonetic database used both for training the models and testing these.

The Section 4 describes the resulting question importance matrix and shows an example of sub tree in more details. Finally some conclusions are given.

2. DECISION TREE CONSTRUCTION

The decision tree construction includes the choice of the:

- The set of the root nodes (it may be just one node for the global decision tree);
- The set of questions, relating to the current frame, for example, the current phone identity, state identity, left or right context identity;
- 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 initial set of nodes (the roots of the decision tree) were the context independent phones, like silence, vowels (different phone models depending on the vowel's attributes like accent and position relating the accent syllable, consonants (the different roots for shifted and non-shifted consonants), the total number of roots were 72 roots.

Each time the tree consists of the tree nodes, each of the nodes corresponds to the HMM for the particular context dependent phone.

The question set includes more than 80 questions for the global tree construction. In the experiments described below the initial roots were chosen to be a context dependent phones with the reduced question set consisting of the following questions:

Is left context is shifted?	(L_Soft ?)?
Is right context is shifted?	(R_Soft ?)
Is left context a lips consonant?	(L_ConLips ?)
Is right context a lips consonant?	(R_ConLips ?)
Is right context a forward vowel?	(R_Forw ?)
Is left contest a forward vowel?	(L_Forw ?)
Is right context a labial?	(R_Labial ?)
Is left context a labial?	(L_Labial ?)

[1] The work was supported in part by Russian Basic Research Fund, grant # 00-01-00648
[2] Computer Center, Russian Academy of Science

Is left context a low vowel?	(L_Low ?)
Is right context a low vowel?	(R_Low ?)
Is left context a mid vowel?	(L_Mid ?)
Is right context a mid vowel?	(R_Mid ?)
Is left context a high vowel?	(L_High ?)
Is right context a high vowel?	(R_High ?)
Is right context a nasal?	(R_Nasal?)
Is left context a nasal?	(L_Nasal?)
Is right context a back consonant?	(R_Back ?)
Is left context a back consonant?	(L_Back?)
Is right context a round vowel?	(R_Round ?)
Is left context a round vowel?	(L_Round?)
Is right context a vowel?	(R_Vow?)
Is left context a vowel?	(L_Vow ?)
Is right context the “m” phone?	(R_M ?)
Is left context the “l” phone?	(L_M ?)
Is right context the “n” phone?	(R_N?)
Is left context the “n” phone?	(L_N ?)
Is right context the “r” phone?	(R_R?)
Is left context the “l” phone?	(L_R?)
Is right context the “r” phone?	(R_Plos?)
Is left context the “l” phone?	(L_Plos?)
Is right context the “j” phone?	(R_J?)
Is left context the “j” phone?	(L_J?)
Does the current frame belong to the first state?	(1ST?)
Does the current frame belong to the second state?	(2ST?)
Does the current frame belong to the third state?	(3ST?)

The decision tree growth algorithm applied is the conventional one and is based on the sequential nodes splitting procedure. Each time moment the couple, consisting of a tree node (hereafter the parent node) and a question from the question list is selected. The question splits the observations belonging to the parent node into the two subsets relevant to the positive and negative implementation of the question. These observation subsets correspond to the couple of the tentative child tree nodes. Then the HMM parameters of these two observation subsets are estimated and two temporary HMM phones are created. For the new HMM models the gain in the log likelihood is calculated in the form of

$$dL(q,n) = L(q,n) + L(!q,n) - L(n),$$

Where L(n) is the estimate of the likelihood of the observations belonging to the parent node n.

L(q,n) – the log likelihood estimate for the observations belonging to the child node of the n-th node that corresponds the positive answer to the question q, L(!q,n) – the log likelihood estimate of the observations belonging to the child node of the node n for the negative answer to the question q.

The couple (q,n) , is selected and the node n will be actually split into the two of child nodes n_a and n_b, if the log likelihood gain dL(q,n) for that couple is the maximum among the all possible couples (q_i,n_i) of nodes $\{n_i\}$ and questions $\{q_j\}$

The process of node splitting (or tree growing) is stopped if the log likelihood gain becomes too small (less than the predefined threshold) or the required number of the phone models is build.

3. SPEECH MATERIAL AND DATA REPRESENTATION

The database used for the design of phone inventory was the bootstrap part of the Russian acoustic-phonetic database for the telephone applications that described in more details in [1]. The database bootstrap part includes approximately 6 hours of manually segmented read speech from phonetically balanced texts of 3060 sentences. There were two synchronous channels recorded in public telephone network and the microphone quality. The microphone quality signal was used in most experiments described below.

The labeling rules of the original database were changed to reduce (via tying) the total number of labels from 91 to 68 labels since some labels (mostly the complex sounds at the word endings) have low occupancy rate in the database.

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). The incoming short time parameters are then coded with four codebooks that in turn represent the 2D self-organized feature map of equal 29*29 dimensions.

Modeling of speech signal for the decision tree creation relies on the discrete left to right HMMs for phone representation, with (for the most cases) 3 states per phone, and no grammar. The final HMM training, re-segmentation and recognition have been performed using Viterbi search procedure.

4. QUESTION IMPORTANCE

The table below describes the 50 first applied questions (independently of the model state) .The target phone described in the leftmost column, the question showed in the middle column, and the total accumulated gain estimation for all frame observations is showed in the right most column.

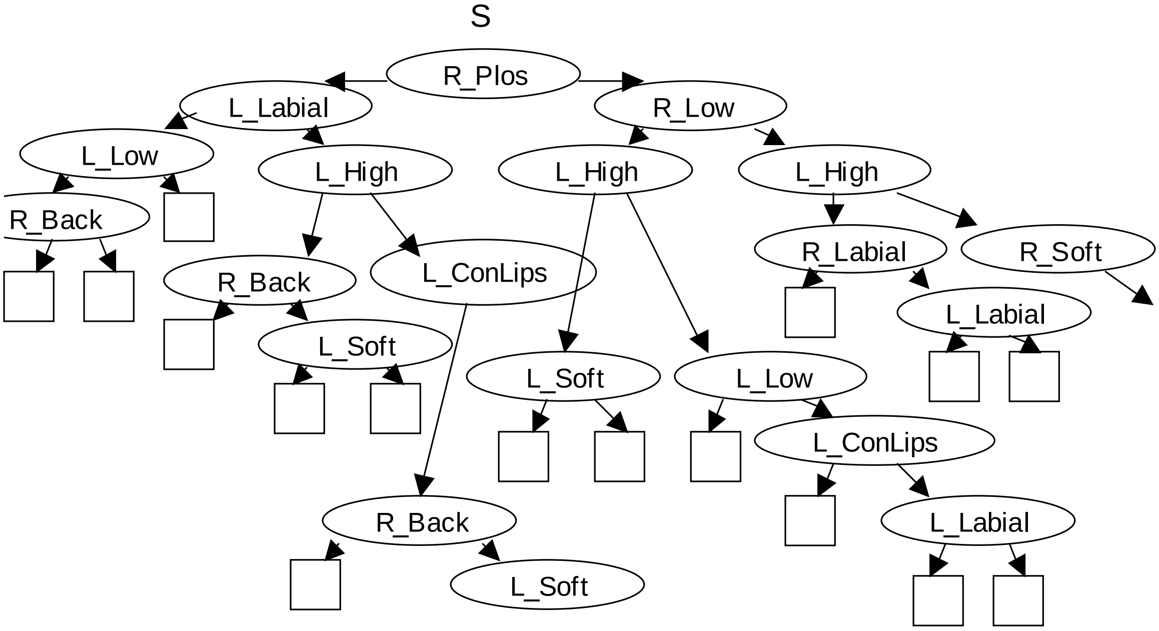
Phone	*Question*	*Total Accumulated Gain*	*Phone*	*Question*	*Total Accumulated Gain*
a	R_Soft	866	je	L_High	72
e^	R_Soft	434	zh	R_Vow	72
a^	R_Soft	382	ts	R_Vow	71
i	R_Soft	322	k'	L_High	70
n	R_Vow	318	b	L_Labial	69
r	R_Vow	276	w0	R_Soft	63
m	R_Vow	244	z'	R_Vow	61
a1	R_Soft	240	n'	R_Forw	60
u	R_Soft	231	y1	R_Soft	59
t	L_Labial	218	ju	R_Plos	57
k	R_Low	216	g	R_Mid	51
o^	R_Soft	213	d'	R_Forw	50
a	R_Plos	200	sh'	L_Soft	48
l'	R_Vow	181	m'	R_Soft	48
s	R_Plos	172	p'	L_Labial	46
w1	R_Soft	169	y	R_Soft	44
l	R_Vow	167	u1	L_Plos	40
j'	R_Vow	157	r'	L_Low	39
s'	R_Vow	154	ja	L_Labial	36
p	L_Low	147	f'	R_High	36
a0	R_Soft	144	b'	L_High	33
i^	R_Soft	142	g'	L_Labial	33
sh	R_Vow	139	v'	R_Mid	32
x	R_Low	129			
ch	R_Vow	129			
v	R_Vow	117			
u^	L_Soft	88			
t'	L_Labial	88			
e	L_Soft	82			
d	R_Vow	82			
y^	L_ConLips	81			
z	R_Low	77			
f	R_Vow	77			

The phone mark symbols corresponds those described in [1]. Prefix “‘” stands for shifted position, prefix “^” stands for stress position, prefixes “1” and “0” denotes the level of a vowel reduction.

The next table depicts the list of 10 most important questions for the all set of phones.

R_Soft	3547
R_Vow	2245
L_Labial	490
R_Plos	429
R_Low	422
L_Soft	218
L_Low	186
L_High	175
R_Forw	110
R_Mid	83

A figure below describes an example of the part of the binary tree for the phone ‘s’.



A part of the decision tree for the phone ‘S’.

5. SUMMARY

There are some conclusions related to the decision tree and importance of the questions that affects on the phone set performance in the recognition tasks.

First, there was a clear dependence between the implementation of the particular question for the particular phone and the number of data observations for that phone. Splitting process tends to produce the data set for all the child nodes to be the same size. This means that if the database used for the design of the phone set has very different phone frequencies from the actual recognition system environment, the resulting decision tree and the designed phone set may be not an optimal one. Also for such a case the attempt to represent “unseen” phones may led to the tying not to the best fitted models.

Second, keeping in mind the first conclusion, the question importance list obtained looks like what was expected from our phonetic knowledge. For example, the right context questions were expected to be generally more important for the vowel parameters than the left context questions, because of the co-articulation effect. Shift consonant influence on the following vowel also was expected to be very important for those vowel acoustic parameters.

LITERATURE

1. Design and implementation of the Russian telephone speech database. In Proc.Int.Workshop “Speech and Computer”, SPECOM’99, Moscow, 1999, p179-181.