Phoneme Set Selection for Russian Speech Recognition

Daria VAZHENINA and Konstantin MARKOV
Human Interface Laboratory
The University of Aizu, Japan
{m5132112,markov}@u-aizu.ac.jp

***Abstract*—In this paper, we describe a method for phoneme set selection based on combination of phonological and statistical information and its application for Russian speech recognition. For Russian language, currently used phoneme sets are mostly rule-based or heuristically derived from the standard SAMPA or IPA phonetic alphabets. However, for some other languages, statistical methods have been found useful for phoneme set optimization. In Russian language, almost all phonemes come in pairs: consonants can be hard or soft and vowels stressed or unstressed. First, we start with a big phoneme set and then gradually reduce it by merging phoneme pairs. Decision, which pair to merge, is based on phonetic pronunciation rules and statistics obtained from confusion matrix of phoneme recognition experiments. Applying this approach to the IPA Russian phonetic set, we first reduced it to 47 phonemes, which were used as initial set in the subsequent speech model training. Based on the phoneme confusion results, we derived several other phoneme sets with different number of phonemes down to 27. Speech recognition experiments using these sets showed that the reduced phoneme sets are better than the initial phoneme set for phoneme recognition and as good for word level speech recognition.**

Speech recognition, Russian language, Phoneme set

I. INTRODUCTION

The first step, we are faced with, in developing any speech recognition system, is choosing appropriate units for acoustic modeling. These units should be accurate, trainable and generalizable. In modern ASR systems for continuous speech, phonemes or phone-like units are usually used. Phoneme set size determines the number of context-independent models, and also influences the number of context-dependent models and the amount of data needed for training. If selected set is too large, the complexity of the phoneme hypotheses lattice will increase significantly, making the decoding process more computationally expensive. If too small, recognition performance may degrade because of low phonetic space resolution.

For Russian language, knowledge based phoneme sets are mostly used. They are manually designed by human experts according to linguistic and phonological rules. The rules for transformation from orthographic text to phonemic representation are not very complicated for Russian, since many words are pronounced the way they are spelled [1]. In [2], 43 phoneme set was used, which consists of the standard SAMPA phoneme set plus additional consonant /h/ because of the data specifics. On the other hand, direct spelling conversion produces 49 phoneme set, which was developed for comparison with grapheme recognizer introduced in [3]. For Russian LVCSR system, 59 phoneme set was proposed in [4], but no results were reported. In most cases, researchers use extended set of vowels including their stressed and unstressed variants [3], [4], [5].

For other languages however, there are studies, where statistical information is utilized for the phoneme set selection. For Chinese language, J.S. Zhang in [6] proposed to use mutual information between the word tokens and their phoneme transcriptions in the training text corpus. Phoneme sets were derived by iteratively merging those tonal-dependent units, which result in minimal mutual information loss. His experiments showed that, when using phoneme set with even slight mutual information loss, number of triphones can be significantly reduced in comparison with the full tone-dependent phoneme set. Word recognition accuracy was also slightly improved.

Statistical methods are also used in designing data-driven subword unit sets. For English language, this kind of phoneme set has been automatically generated given a set of acoustic signals and their transcriptions [7]. A drawback of this approach is that it is difficult to add new words to the speech recognition system dictionary.

In this study, we select phoneme sets using both phonological knowledge and statistical information. Applying this approach to the IPA Russian phonetic set, we first reduced it to 47 phonemes, which were used as initial set in the subsequent speech model training. In Russian, almost all phonemes come in pairs: hard and soft consonants and stressed and unstressed vowels. These pairs are considered as candidates for further merging. After a phoneme recognition experiment, based on the information from the confusion matrix, we calculate confusion rates for those pairs. Pairs with highest confusion rates are merged. When new phoneme set is obtained, speech recognition lexicon is updated appropriately. For evaluation, we performed phoneme recognition using monophone models as well as word recognition using monophone and cross-word triphone models.

II. RUSSIAN PHONOLOGY SPECIFICS

The International Phonetic Alphabet (IPA) is often used as practical standard phoneme set for many languages. For Russian it includes 55 phonemes: 38 consonants and 17 vowels [8]. The large number of consonants is caused by the specific palatalization in the Russian language. All but eight of the

978-1-61284-729-0/11/$26.00 ©2011 IEEE

consonants occur in two varieties: plain and palatalized. For example, hard /b/ and soft /b'/, as in 'небо' (sky) and 'берёза' (birch). This is caused by the letter that follows the consonant and appears as secondary articulation by which the body of the tongue is raised toward the hard palate and the alveolar ridge during the articulation of the consonant. Such pairs make speech recognition task more difficult, because they increase consonant confusability in addition to the fact that they are less stable than vowels and have smaller duration. In comparison, IPA set for American English includes 49 phonemes: 24 consonants and 25 vowels and diphthongs.

Russian IPA vowel set includes also phoneme variants with reduced duration and reduced articulation. There are six base vowels in Russian phonology [9], that are often used as stressed and unstressed pair, for example /a!/ and /a/. In unstressed syllables, all of them are subject to vowel reduction in duration and all but /u/ to articulatory vowel reduction tending to be centralized and becoming schwa-like [10]. Thus, unstressed /e/ may become more central and closer to unstressed /i/, and unstressed vowel 'o' is almost always pronounced as /a/ except in case of foreign words such as 'радио' (radio).

III. Description of the Selection Method

In our approach, we use information from phonological knowledge and statistics from the confusion matrix of phoneme recognition experiment. Phoneme set selection workflow, shown on Fig. 1, includes following steps:

1. First, from the IPA set, we derive phoneme set P0 by applying phonological.
2. For further merging, we define phoneme pair candidates according to language phonology specifics. Those include both soft and hard consonant and stressed and unstressed vowel pairs.
3. Using P0 set, we perform phoneme recognition and obtain phoneme confusion matrix. For selected pairs, we calculate confusion rate (CR) as follows:

$$CR = \frac{M_1 + M_2}{H_1 + M_1 + H_2 + M_2} * 100\% \qquad (1)$$

here H_1 is number of correctly recognized occurrences of the first phoneme in the pair, e.g. /a/ recognized as /a/, H_2 is number of correctly urrences of the second phoneme in the pair, e.g. /a!/ recognized as /a!/, M_1 is number of misrecognized occurrences of the first phoneme in the pair, e.g. /a/ recognized as /a!/, and M_2 is number of misrecognized occurrences of the second phoneme in the pair, e.g. /a!/ recognized as /a/. The higher confusion rate, the more phonemes are mismatched within the pair, which makes it a likely candidate for merging.

4. Phoneme pairs are sorted by decreasing confusion rates.
5. Finally, we select top N phoneme pairs and by merging them obtain new phoneme set. Different values of N produce different phoneme sets.

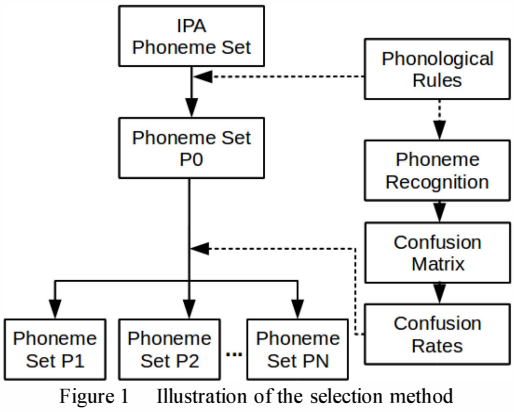

Figure 1 Illustration of the selection method

The best phoneme set can be found by evaluating its performance in speech recognition. Analyzing the results, we can make final decision about phoneme set to choose.

IV. Experiments and Results

A. Databases and Feature Extraction

In our experiments, we used SPIIRAS [11] and GlobalPhone [3] Russian speech databases. Speech data are collected in clean acoustic conditions. SPIIRAS database consists of 16350 utterances pronounced by 50 speakers (25 male and 25 female) with total duration of about 21 hours. Speech recognition results were obtained after 5-fold cross validation. In each fold, utterances from different 5 male and 5 female speakers were used for testing. GlobalPhone database consists of 12321 utterances pronounced by 115 speakers (61 male and 54 female) with total duration of about 26 hours. Data were split to training set - 90\% and test set - 10\% of the database. There are 5 male and 5 female speakers in the test data.

We used the HTK toolkit [12] to build our speech recognition systems. Acoustic signal was coded with energy and 12 MFCCs (Mel Frequency Cepstral Coefficients) and their first and second derivatives, resulting in 39-dimension feature vector. Acoustic phoneme models were represented by three state HMMs with left-to-right topology except the silence model, which also has transition from third to the first state. Each state pdf was modeled with 16 component Gaussian mixture. Triphones were clustered by phonetic decision tree state tying using custom question set.

Speech corpus transcriptions are used as training data for language modeling. In SPIIRAS database, there are 323 unique sentences with 2332 words. Language model is built as closed set back-off bigram model with perplexity of 241. The lexicon size consists of 1356 entries including 1146 unique words and 210 pronunciation variants. In GlobalPhone database, there are 6456 unique sentences with 106541 words. Language model is built in the same way and has perplexity of 137. The lexicon size consists of 22027 entries including 19973 unique words and 2054 pronunciation variants.

B. Phoneme Set Selection Experiment

To obtain phoneme set P0 we merged stressed vowels with the corresponding closest acoustic neighbor, for example 'a' and

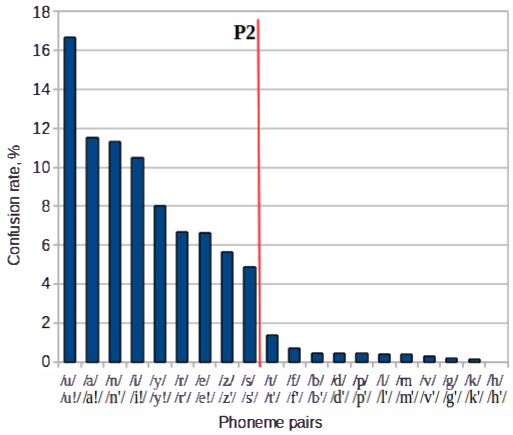

Figure 2 Confusion rates for all phoneme pairs sorted in descending order using SPIIRAS database. P2 shows the N-best threshold for phoneme set P2.

'ɑ'. In addition, consonants 'ʑ' and 'ɣ' were excluded, because they are used only in some dialects in conversational speech. In total, P0 set includes 47 phonemes: 6 stressed and 5 unstressed vowels and 36 consonants, same as in SAMPA alphabet (see Table I for details). As candidate phoneme pairs for merging, hard and soft consonants were selected, because, according to Russian phonology, they have close articulatory positions. In addition, stressed and unstressed vowel pairs were chosen because their main difference is in duration. Vowel reduction in articulation is already embedded in the pronunciation lexicon word forms.

TABLE I PHONEME SET P0.

Consonants	Hard	b, v, g, d, zh, z, k, l, m, n, p, r, s, t, f, h, c, sh
	Soft	b', v', g', d', z', j, k', l', m', n', p', r', s', t', f', h', ch, sch
Vowels	Stressed	a!, e!, i!, o!, u!, y!
	Unstressed	a, e, i, u, y

Using set P0, we built two phoneme recognition systems and performed phoneme recognition experiments for both databases separately. From the resulting confusion matrices we make a list of all candidate phoneme pairs sorted by their confusion rates. This is shown in Fig. 2 for SPIIRAS database and in Fig. 3 for GlobalPhone database.

As can be seen, the most confusable pairs in both systems

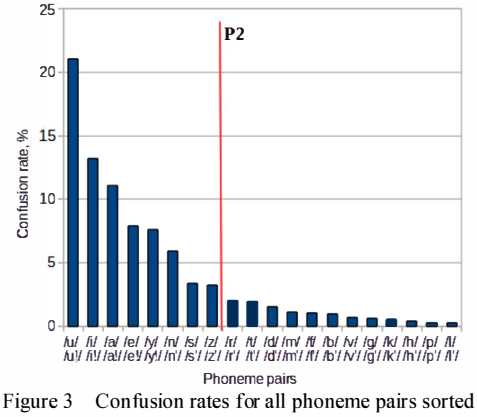

Figure 3 Confusion rates for all phoneme pairs sorted in descending order using GlobalPhone database. P2 shows the N-best threshold for phoneme set P2.

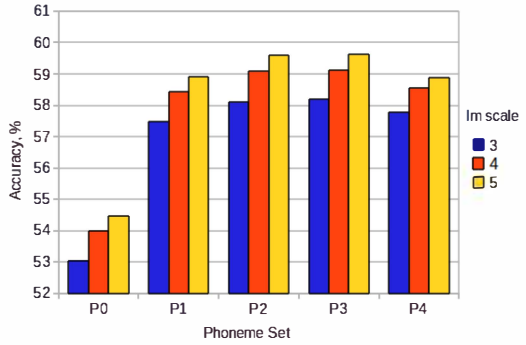

Figure 4 Phoneme recognition results using monophone models for SPIIRAS database.

are the vowel pairs. Thus, we first merge all vowel pairs and obtain phoneme set P1.

Next we look at the consonant pairs and see, that top ten pairs in both systems are very similar, slightly differing in order. We put N-best thresholds between pairs with quite big difference in confusion rates.

Finally, we merged all candidate pairs in P4. Summary of all phoneme sets is given in Table II.

TABLE II NUMBER OF SELECTED PHONEME SETS USED IN EXPERIMENTS.

Phoneme set	SPIIRAS		GlobalPhone	
	Phonemes number	Remarks	Phonemes number	Remarks
P0	47			
P1	42	P0 without /a!/, /e!/, /i!/, /u!/, /y!/		
P2	38	P1 without /n'/, /r'/, /s'/, /z'/	39	P1 without /n'/, /s'/, /z'/
P3	37	P2 without /t'/	37	P2 without /r'/, /t'/
P4	27	P3 without /b'/, /g'/, /d'/, /k'/, /l'/, /m'/, /p'/, /f'/, /h'/		

C. *Phoneme set evaluation*

For both databases, in addition to the system based on P0, we trained 4 more systems using phoneme sets P1-P4. Phoneme recognition performance of these systems using monophone models is shown in Fig. 4 for SPIIRAS database, in Fig. 3 for GlobalPhone database, for three different language model (lm) scales. The language model is simple phoneme bigram trained on the phoneme transcription of the data. Phoneme recognition results showed similar trends in changing of accuracy for both databases. Recognition accuracy for P2 and P3 phoneme sets is almost equal and higher than the one

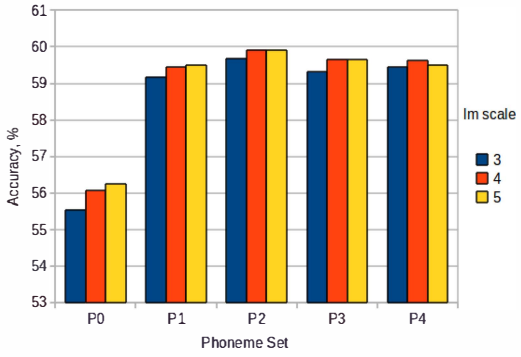

Figure 5 Phoneme recognition results using monophone models for GlobalPhone database.

for P0 and P4. Difference between results of P0 phoneme set and all others can be explained by the absence of highly confusable vowel pairs. Also, results for the P2 and P3 phoneme sets are higher than for P1, because the mismatch of several consonant pairs was reduced by their merging. Slightly worse performance of P4 suggests that too many consonants have been merged resulting in decreased phonetic space resolution. This phoneme recognition experiment showed that candidate pairs were chosen correctly and their merging gives better results.

TABLE III NUMBER OF TRIPHONES AND STATE POOL SIZES FOR DIFFERENT PHONEME SETS.

Phoneme set	SPIIRAS		GlobalPhone	
	# Triphones	# States	# Triphones	# States
P0	112849	2621	112849	3096
P1	81314	2635	81314	3118
P2	60802	2585	65562	3178
P3	56279	2555	56279	3170
P4	22709	2542	22709	3111

Next, word recognition experiments were performed using cross-word triphone models. Number of triphones for each phoneme set is shown in Table III for both systems. It also shows the number of tied states, which we tried to make similar during systems development. Word recognition results, shown on Fig. 6 for SPIIRAS database, although higher are similar to ones shown on Fig. 7 for GlobalPhone database.

Word recognition results showed that:

- Merging of all phoneme pairs as in P4 gives worse results, which can be due to reduced resolution in the phonetic space.

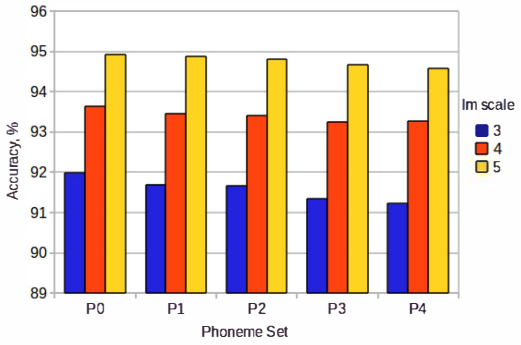

Figure 6 Word recognition results using triphone models for SPIIRAS database.

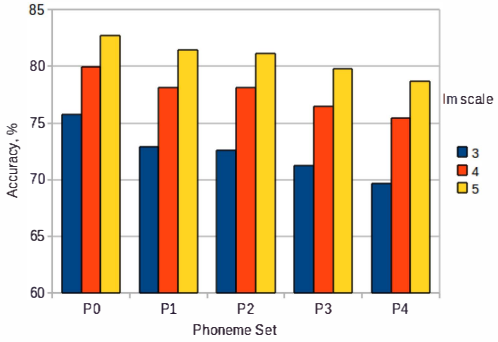

Figure 7 Word recognition results using triphone models for GlobalPhone database.

- Difference in accuracy between initial phoneme set P0 and P1, P2 is very small, but difference in triphone number is considerable.

V. CONCLUSION

In this paper, we presented a phoneme set selection method for Russian language, where statistical information is used along with linguistic and phonological knowledge about the language. Applying our method to the SPIIRAS and GlobalPhone speech corpora, similar results were obtained. The top list of most confusable phoneme pairs slightly differs in pairs' order, while the consistency of this list is the same. Speech recognition results for triphone models showed similar trend of changing word recognition accuracy depending on selected phoneme set. Phoneme recognition accuracy is higher for the phoneme set with frequently mismatched vowel and consonant pairs.

The phoneme recognition experiments demonstrate the potential of acoustic models for matching phonemes better, when the most confusable phonemes are merged. In the further research, the optimal stop criterion for phoneme set selection may be investigated. Possibly, speech recognition experiments using more complicated language model may help to obtain this criterion. Language model could be changed by using higher order n-grams with larger number of unique words.

REFERENCES

[1] Paul Cubberley, Russian: a linguistic introduction, Cambridge University Press, 2002.

[2] Josef Psutka, Pavel Ircing, J.V. Psutka, Jan Hajic, William J. Byrne and Jiri Mirovsky, "Automatic transcription of Czech, Russian, and Slovak spontaneous speech in the MALACH project," in Proc. INTERSPEECH, Lisbon, Portugal, September 2005, pp. 1349–1352.

[3] Sebastian Stuker and Tanja Schultz, "A grapheme based speech recognition system for Russian," in Proc. SPECOM, St.Peterburg, Russia, Sep 2004, pp. 297–303.

[4] Marina Tatarnikova, Ivan Tampel, Ilya Oparin, and Yuri Khokhlov, "Building acoustic models for a large vocabulary continuous speech recognizer for Russian," in Proc. SPECOM, St.Peterburg, Russia, June 2006, pp. 83–87.

[5] Andrey Ronzhin and Alexey Karpov, "Automatic system for russian speech recognition SIRIUS," in Proc. SPECOM, St.Peterburg, Russia, September 2004, pp. 291–296.

[6] Jin-Song Zhang, Xin-Hui Hu, and Satoshi Nakamura, "Using mutual information criterion to design an efficient phoneme set for chinese speech recognition," IEICE Transactions on Information and Systems, vol. E91-D, no. 3, pp. 508–513, March 2008.

[7] Rita Singh, Bhiksha Raj, and Richard M. Stern, "Automatic generation of subword units for speech recognition systems," IEEE Transactions on Acoustics, Speech and Signal Processing, vol. 10, no. 2, pp. 89–99, February 2002.

[8] "IPA international phonetic alphabet," http://en.wikipedia.org/wiki/Wikipedia:IPA for Russian.

[9] "SAMPA computer readable phonetic alphabet," http://www.phon.ucl.ac.uk/home/sampa/russian.htm.

[10] Jaye Padgett and Marija Tabain, "Adaptive dispersion theory and phonological vowel reduction in russian," Phonetica, vol. 62, no. 1, pp. 14–54, 2005.

[11] Oliver Jokisch, Agnieszka Wagner, Robert Sabo, Rainer Jaeckel, Natalia Cylwik, Milan Rusko, Andrey Ronzhin, and Ruediger Hoffmann, "Multilingual speech data collection for the assessment of pronunciation and prosody in a language learning system," in Proc. SPECOM, St.Peterburg, Russia, June 2009, pp. 515–520.

[12] Steve Young et al., The HTK Book, Cambridge Univ., 2009.