Building Acoustic Models for a Large Vocabulary Continuous Speech Recognizer for Russian

Marina Tatarnikova, Ivan Tampel, Ilya Oparin, Yuri Khokhlov,

Speech Technology Center
St. Petersburg, Russia
{tatmar,tampel,ilya}@speechpro.com

Abstract

Different types of acoustic models created at Speech Technology Center are evaluated in this paper. Our main goal was to test how well those models work and choose one model for implementation in a large vocabulary continuous speech recognition (LVCSR) system for Russian which is under development now. Context-independent discrete and continuous models, as well as context-dependent continuous models, were built and evaluated on an isolated word recognition task. The results gained with the context-dependent continuous model prove its consistency and show it can be used for acoustic modelling in a large vocabulary speech recognizer.

1. Introduction

Any modern LVCSR system can be roughly divided into three major parts: the acoustic model, the language model and the decoder. Such a division is inferred from the main formula of speech recognition. This is the formula for computing the probability that a certain linguistic unit *W* corresponds to a given acoustic signal *O*:

$$P(W \mid O) \tag{1}$$

It is not possible to calculate this probability directly and Bayes rule is used to rewrite it as

$$P(W \mid O) = \frac{P(O \mid W)P(W)}{P(O)} \tag{2}$$

Since the probability of the acoustic signal itself *P(O)* is constant for all recognition candidates and we are looking for the maximum score among them, this component can be dropped out. Finally, the task is reduced to calculating

$$P(O \mid W)P(W) \tag{3}$$

P(O|W) is called *acoustic likelihood* (since it tells us how likely it is that the speech segment corresponds to some particular linguistic unit) and *P(W)* is the so-called *prior* probability (it corresponds to the likelihood of the linguistic unit itself, as it is in the language). The former is calculated by the acoustic component of a speech recognizer, while the calculation of the latter pertains to the language modelling component.

Usually the basic unit of a language model is the word and the total score is obtained as the best path in recognition network which combines all possible word sequences together with individual acoustic likelihoods and priors for all possible units (words). Due to a huge search space in the case of LVCSR, finding the optimal path is a complicated task which calls for implementation of advanced network search techniques [1]. This is done by a special part of a speech recognizer conventionally called a *decoder*.

At present one can claim there is no decently working LVCSR system for Russian. The work on a LVCSR system for the Russian language is being carried out at Speech Technology Center (`http://www.speechpro.com`). All the components mentioned above are being built from scratch using our own technologies. Those technologies are adapted to the peculiarities of the Russian language.

This paper is focused on the acoustic model of the automatic speech recognition (ASR) system. However, other components are also briefly sketched in chapter 7 in order to provide the reader with an insight into the state-of-the-art in the ASR for Russian.

At present most of ASR systems are based on the Hidden Markov Model (HMM) approach (see [2] for details). HMM is a powerful statistical approach which represents speech as a parameterized random process. Each modelled speech object (word, syllable, phoneme etc.) is represented by its own HMM.

Previously at Speech Technology Center (STC) we used acoustic models based on HMMs for command recognition (i.e. small vocabulary isolated word recognition). Individual acoustic models were generated for whole words (commands) as homogeneous units. Good results of isolated word recognition were gained using this technique. However, this approach has insurmountable limitations which keep it within the command recognition domain only. It is virtually impossible to create such whole-word acoustic models in a large-vocabulary system. For modern LVCSR systems common size of the vocabulary is 65K words, but it may be extended up to several hundred thousand. That means acoustic models should be constructed for smaller units and then concatenated to model larger ones. Syllables, phonemes and even phoneme fragments were tested as such units.

At present, context-independent phonemes (monophones) for medium size vocabulary recognition and context-dependent phonemes (diphones and triphones) for LVCSR are used. This is because more precise acoustic description is needed to distinguish between thousands of words. In general, diphones and triphones are used to compensate for the effect of coarticulation in continuous speech. Coarticulation is the result of the fact that articulatory organs never get the static positions as for isolated sounds but rather show the movement in the needed direction. This direction depends both on preceding and succeeding phones. Coarticulation is not limited only to the neighbouring phones but may involve several phones in the context. However, only immediate

context is taken account of in case of diphones (only left neighbour) or triphones (both left and right neighbours). On the analogy, models which do not take account of context are called monophonic.

The general structure of the paper is as follows. Issues regarding processing of audio signal and techniques of feature extraction are discussed in Chapter 2. Chapter 3 is devoted to the peculiarities of acoustic models structure and feature estimation. Results for different types of acoustic models and the evaluation of these results are presented in Chapters 4 and 6 correspondingly. Chapter 5 shows the performance of the automatic monophone segmentation performed with the acoustic models described in this paper. Chapter 7 gives a short overview of the development of the speech recognizer on the whole.

2. Speech Signal Processing and Feature Extraction

Mel-frequency cepstral coefficients (MFCC) were chosen for preliminary processing of speech signal. MFCC are widely used in the field of ASR [3].

We also developed a novel method of feature extraction by means of a special filter bank that consists of recursive filters of the second order. This method appeared to be more robust to high noise level in speech signals. When compared to MFCC, recognition results gained with our method appeared a bit worse for clean speech signals. However, for noisy signals, we get better performance starting from 15 dB level of signal/noise ratio. If this ratio is further decreased (i.e. the noise level grows up) the profit of using our method is further increased.

At present, robustness to noise is one of the main issues in the field of ASR. It is very important to keep system performance stable even in the case heavily distorted speech is being recognized. The method sketched above (its thorough description calls for a separate paper) is implemented in the LVCSR system we are developing. However, MFCC coefficient approach was chosen in this paper just because the latter is conventionally used for the comparison of results between different systems.

Input signal is quantized at 11025 Hz and transformed into a set of feature vectors. The signal is analyzed within a 256-sample window with a 128-sample step.

The signal is pre-emphasized in a usual way by a first order FIR filter:

$$S'(n) = S(n) - k * S(n-1) \tag{4}$$

where $n \in \{1 \ldots N\}$;

N – window dimension;
$S(n)$ – signal samples;
k – amplification coefficient.

Hamming window is used for weakening signal distortion caused by the application of discrete window to a continuous signal.

Human ear is known to perceive frequencies non-linearly in the audio spectrum [4]. As a result, an analyzer performing preprocessing of speech signal with a non-linear scale boosts the recognition rate. We used Mel-scale frequency filters as a rather conventional and wide-spread solution for this task. Amplitudes obtained from the set of triangle filters are highly correlated. Cosine transform (used to obtain cepstrum) was implemented for feature decorrelation:

$$C_i = \sqrt{\frac{2}{N}} \sum_{j=1}^{N} m_j \cos\left(\frac{PI * j}{N}(i - 0.5)\right) \tag{5}$$

where $i \in \{1 \ldots M\}$;

$M=12$ – number of cepstral coefficients;
$N=20$ – number of filters;
m_j – logarithmic amplitudes.

In addition to 12 cepstral coefficients, the value of energy was added. It is calculated within the analysis window:

$$E = \log \sum_{n=1}^{N} s_n^2 \tag{6}$$

where s_n – speech samples before processing;
N – length of the analysis window.

Precision of the recognition is increased if the basic parameters described above are supplemented with the first and the second time derivatives. These derivatives correspond to the speed and acceleration parameters (Δ and ΔΔ coefficients). Those are calculated for all 12 MFCC and for the energy. The dimension of the resulting feature vector is 39 and it consists of 4 groups of features:

- Energy, energy Δ and energy ΔΔ;
- 12 cepstral coefficients;
- 12 Δ-features;
- 12 ΔΔ-features.

3. Acoustic Models

3.1. Model Parameters

The acoustic model of a recognition system is a set of HMM events. Such events in acoustic modelling are usually allophones (monophones, diphones or triphones). Markov model $\lambda(A, B, \pi)$ of an acoustic event is one or several states, which are characterized by the following parameters:
N – number of states;
π – initial distribution of probabilities;
A – transition (from one state to another) matrix;
B – probabilistic density function in the feature space (so-called *emission probability*).

For discrete HMMs a codebook of feature space and the probabilistic density function is created. The latter is represented as a matrix of codewords probabilities for a given state $B_i(k)$, where k is the word index in the codebook and i is the state number.

In the case of continuous models probability density function B is represented by means of a combination of M Gaussian functions:

$$B_i(x) = \sum_{m-1}^{M} C_{im} G[x, \mu_m, U_m] \tag{7}$$

where x – observation vector;

c_{im} – weight coefficient (contribution of a Gaussian component *m* to probability density function for the i[th] state);
μ_m – mean value of the feature vector for a Gaussian mixture component *m*;
U_m – covariation matrix for the m[th] feature component;
G – multidimensional Gaussian function

$$G[x,\mu_m,U_m]=\frac{1}{\sqrt{(2\pi)^n|U|}}\exp(-\frac{1}{2}(x-\mu)^T U^{-1}(x-\mu) \quad (8)$$

where *n* – feature vector dimension;
$|U|$ – determinant of the covariation matrix U.

Diagonal covariation matrixes were used in our study. On the whole, This approach is known as Gaussian Mixture Model (GMM).

3.2. Model Topology

The inhomogeneous left-to-right monophone model was used for acoustic modelling in the system presented in this paper:

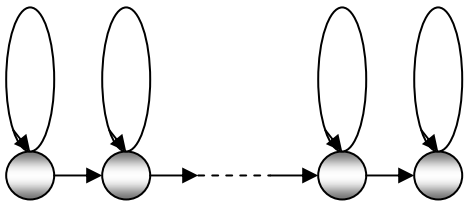

Figure 1: Acoustic Model Topology

Each monophone is representd by 1-3 states, depending on its duration and its representation in the training corpus. State transition probabilities are not constant and depend on time already spent in the current state. Such a model is known as an *Inhomogeneous Markov Model* [5]. Jumps across the states are not allowed by the model topology. This way temporal state features contribute most to the formation of the accumulated probability. The extent of influence of state durations is optimized on the basis of recognition results. This is done by increasing or decreasing mean square deviations for histograms describing life time of the states.

3.3. Model Types

Manually segmented part of the database was used for the initial calculation of Gaussians that approximate probability density distributions in the feature space.

At the beginning of the training procedure each state is described by only one Gaussian. Increase in the number of Gaussian functions per state is controlled by a threshold function based on the entropy measure. Since each acoustic feature vector consists of four groups of parameters, different sets of Gaussians were built for each group separately.

Three types of models were tested for phoneme recognition task:

1. **Context-Independent discrete models.** A word model is obtained by the concatenation of the monophone HMMs. The same monophones are described by one model. The number of monophones is equal to 40 in our case. Emission probabilities are calculated on the basis of the codebook.
2. **Context-Independent continuous models.** Models of this type share all the properties with the Context-Independent discrete models except for the fact that GMMs were used for the approximation of probability density functions (while discrete models use the codebook approach). A unique set of Gaussian functions is created for each model.
3. **Context-Dependent continuous models.** A word model is obtained by the concatenation of the monophone models. However, those monophones are not estimated on the basis of all idem monophones in the training corpora (i.e. coming from other words). Certain monophones from the same word are only used for training. For example, when building the model for the word *дом*, each constituent monophone model is estimated not on the basis of all instances of monophones '*д*', '*о*', '*м*' in different words (as in the context-independent models), but only on the basis of those monophones which occurred in different instances of the word *дом* in the training corpus. This way the context is taken into account. Each state model is described by its own set of Gaussian functions.

4. Evaluation

4.1. Training and Evaluation Corpora

All the algorithms of the creation of acoustic models for monophones were trained and evaluated on the acoustic database of Speech Technology Center. It consists of recordings of 62 speakers (30 male and 32 female). The recognition dictionary for this database consists of 23 commands. The commands included 10 digits from 0 to 9 and 13 commands used to operate a cellular phone hands-free device. The commands were pronounced in an isolated manner. Speech was recorded in a car with a microphone used in a hands-free device and sampled at 11025 Hz. Car engine was idling during the process of recording.

During the evaluation procedure 7 groups with 6 speakers in each were being successively excluded from the training corpus and used as the evaluation set. The rest of the corpus was used for training.

4.2. Results

Recognition results are presented in per cents in Table 1. The letters in the first column correspond to different groups of speakers mentioned in the previous section.

The results obtained for the phoneme-based models described above are compared against the results for the whole-word recognition method (shown in the second column). In the case of whole-word recognition HMMs were constructed not for the constituent phonemes but for the whole words (i.e. pronunciation variants). The number of HMM states in this case is variable, depending on the number of phones in the word. Feature space was quantized. Probability density function for each state in the feature space is represented by a matrix of probability of codeword

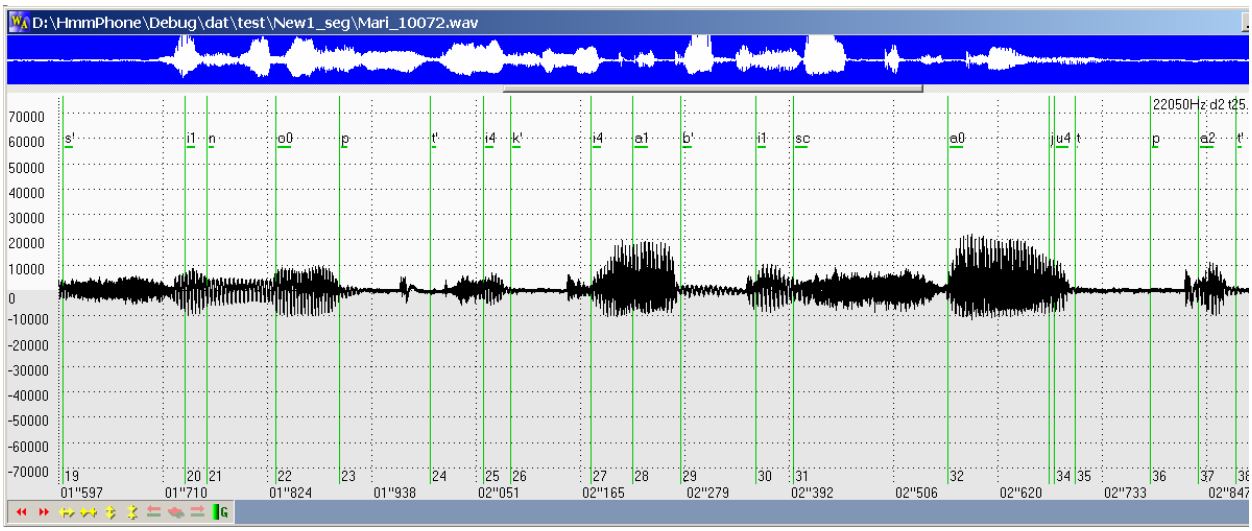

Figure 2: Automatic Segmentation of Speech in Monophones

occurrence in a given state. This approach has critical limitations described in Chapter 1. However, it secures good results for command recognition by default and can serve as a baseline in our evaluation.

Table 1: Command Recognition Results

	Whole-word model	Context-Independent discrete model	Context-Independent continuous model	Context-Dependent continuous model
A	98,36	93,99	98,36	100,0
E	98,07	95,37	97,68	100,0
G	98,12	94,84	98,12	100,0
K	97,93	95,45	97,93	99,59
L	100,0	97,13	97,54	100,0
M	97,93	95,52	97,93	98,97
N	98,37	95,42	98,04	98,69
All	98,4 (ERR=1,6)	95,39 (ERR=4,6)	97,94 (ERR=2,0)	99,6 (ERR=0,4)

We also tested dynamic programming recognition methods on the same test material. The error rate of the whole-word speaker-independent ASR based on dynamic programming appeared at the level of 4%, which is significantly worse than for any of the continuous HMM-based models described above.

5. Continuous Speech Recognition

The algorithms discussed in this paper were used in the LVCSR system which is currently under development. At first context-independent continuous models were chosen for building acoustic models. We are planning to switch to the triphone context-dependent models only when models of independent monophones are fully implemented. That is due to the fact trees of triphones (that take account of context) are to be build on the basis of the monophonic models.

Acoustic models for Russian speech were built for 59 independent monophones. We use 3 Gaussian functions for modelling energy features for each state, while features of other 3 groups discussed in chapter 2 are modelled with 6 Gaussians on average. A phonetically balanced text read by 203 speakers (111 male and 92 female) was used as a training corpus. Speech was recorded with a high quality microphone Philips SBC MD110, Sound Blaster Live! and was sampled at 22050 Hz.

The speech of 7 speakers was manually segmented and used at the initial phase of training. The rest of the corpus was segmented automatically on the basis of estimations obtained at the initial phase. An example of the automatic segmentation is presented in Figure 2. The segmentation of the fragment of a phrase “На следующей неделе синоптики обещают потепление” (Warming is forecast next week) is shown on the waveform.

We are not able to perform full-scale ASR evaluation yet, since not all components of the LVCSR system are ready. However, results of automatic segmentation can be considered as a tool for preliminary and rough evaluation of the acoustic model as a part of the recognition system. Automatic segmentation is performed by means of acoustic models. Those are initially trained on the basis of hand-labeled part of the training corpus. Acoustic models obtained this are undertrained. However, it is very costly to segment manually a larger part of the corpus that is required for acoustic model training. That is why “initial” acoustic models are used to segment the rest of the corpus for which only the transcription (but not the positions of monophone boundaries) are known. Newly obtained segmented data is used for the re-estimation of acoustic model parameters. That means the quality of automatic segmentation reflects the quality of acoustic models themselves.

In out case, the quality of automatic monophone segmentation appears on the high level. As can be seen from Figure 1, most monophone boundaries are detected with high precision. However, this approach does not let evaluate the performance numerically, which is obviously its main shortcoming.

6. Conclusions

Recognition results obtained for different phone models turned out to be on the same level as for an ad hoc whole-word model. This tells us that the acoustic models we are planning to use in an LVCSR system are consistent and can be successfully implemented.

As expected, the best recognition results were gained with context-dependent continuous models. Discrete models led to the 3 times increase in the error rate. That means using coherent GMMs for the estimation of the probability density function is highly beneficial.

We claim the internal HMM structure we proposed is also coherent. This is proved by the comparison of the results gained for HMM and dynamic programming based models which show significant advantage of the former. Another proof is the high quality automatic monophone segmentation gained with the acoustic models we use.

7. Parallel Work

Our next step is to build context-dependent continuous model into the general frame of a LVCSR system. To be precise, we imply triphone models (mentioned in chapter 5) but not the ones described in section 3.3 as context-dependent models. The former are less context-dependend since they take account only of an immediate left and right context of a phone, while the latter were bound to a whole word.

Building acoustic models into an LVCSR framework makes sense only if other parts (i.e. the language model and the decoder) of a recognition system are ready. At present those are still under construction, however, most of the work has already been completed.

A tool for language modelling was initially developed at STC last year. This tool allows building two types of model: standard wordform-based and stem/inflexion ones. The necessity of introduction of morphological information into a language model results from the peculiarities of Russian as an inflective language. This feature of the language gives rise to a whole bunch of problems which are not that crucial for languages with a weak inflexional system (and relatively strict word order), like English. Most significant is the high OOV rate and impetuous growth of wordform-based pronunciation vocabulary when one tries to lower the size of the vocabulary down [6, 7]. Maintaining a huge vocabulary can be troublesome. Many decoders imply limitations on the vocabulary size (however, this is not the case for the decoder developed at STC). But much more important is that the quality of a pronunciation vocabulary can greatly influence the recognition rate, but at the same time it is hardly possible to perform manual checking and updating of a huge vocabulary. Even if such checking is performed only for a restricted set of words in the vocabulary, additional problems arise. For example, to change transcription in the root of a word, one would have to make the same changes in many other wordforms of this word. At the same time automatic approach is not expected to be easily implemented, since changes could be applied only to a subset of wordforms. In case a new pronunciation is to be added for some lemma, even more effort is required. That is why morphemic models are often suggested as a solution [8, 9, 10] to this problem. Morphological information allows keeping pronunciation vocabularies compact; another feature is that less training data is needed to train a morphologically-based LM. The morphological system we implemented is described in [11]. Now the language modelling tool is implemented as an SQL-based application which makes the LM easy and fast to handle.

The decoder we use is a time-synchronous one-pass stack decoder. It allows handling huge vocabularies (hundreds of thousands of units) which makes it applicable to the wordform-based language model of the Russian language. Details regarding the decoder structure and its implementation can be found in [12].

8. References

[1] X.L. Aubert, "An overview of decoding techniques for large vocabulary continuous speech recognition", *Computer, Speech and Language, (16) 2002. pp. 89-114.*

[2] F. Jelinek, "Statistical Methods for Speech Recognition", *MIT Press, Cambridge, Massachusetts, 1997.*

[3] S. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev, P. Woodland. The HTK Book, http://htk.eng.cam.ac.uk

[4] H. Hermansky, N. Morgan, "RASTA Processing of speech", *IEEE Trans. On Speech and Audio Processing, V.2, N.4, October 1994. pp.587-589.*

[5] P. Ramesh, J.G. Wilpon, "Modeling state durations in hidden Markov models for automatic speech recognition", *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing, 1992. pp. 381–384.*

[6] Whittaker E.W.D., Woodland P.C., "Comparison of Language Modelling Techniques for Russian and English", *Proc. of ICSLP'98, Twente, the Netherlands, 1998.*

[7] Siivola V., Kurimo M. and Lagus K., "Large Vocabulary Statistical Language Modeling for Continuous Speech Recognition in Finnish", *Proc of Eurospeech'01, Denmark, Aalborg, 2001, pp 737-740.*

[8] W. Byrne, F. Jelinek, P. Ircing, P. Krbec, J. Hajic, J. Psutka, S. Khudanpur, "On Large Vocabulary Continuous Speech Recognition of Highly Inflectional Language – Czech", *Proc. of 7th Eurospeech'01, Denmark, Aalborg, 2001, pp. 487-490.*

[9] V. Siivola, T. Hirsimaki, M. Creutz and M. Kurimo, "Unlimited Vocabulary Speech Recognition Based on Morphs Discovered in an Unsupervised Method", *Proc. of Eurospeech'03, Geneva, Switzerland, 2003. pp. 2293-2296.*

[10] A.L. Ronzhin, A.A. Karpov, "Implementation of morphemic analysis for Russian speech recognition", *Proc. of SPECOM'2004, St. Petersburg, Russia, 2004, pp. 291-296.*

[11] I. Oparin, A. Talanov, "Stem-Based Approach to Pronunciation Vocabulary Construction and Language Modeling of Russian", *Proc. of the 10th International conference on Speech and Computer, SPECOM 2005. Patras, Greece, 2005. pp. 575-578.*

[12] M. Zhenilo, A. Ivanov, I. Oparin, "Efficient Linear Recursive Hashing Method for Look-Ahead of the n-Gram LM with Back-Off", *Submitted to the 11th International conference on Speech and Computer, SPECOM 2006. St.Petersburg, Russia, 2005.*