INFORMATICA, 2004, Vol. 15, No. 4, 465–474

Development of HMM/Neural Network-Based Medium-Vocabulary Isolated-Word Lithuanian Speech Recognition System

Mark FILIPOVIČ, Antanas LIPEIKA

Recognition Processes Department, Institute of Mathematics and Informatics
Goštauto 12-204, LT-01108 Vilnius, Lithuania
e-mail: markas@mch.mii.lt, lipeika@ktl.mii.lt

Received: December 2004

Abstract. The development of Lithuanian HMM/ANN speech recognition system, which combines artificial neural networks (ANNs) and hidden Markov models (HMMs), is described in this paper. A hybrid HMM/ANN architecture was applied in the system. In this architecture, a fully connected three-layer neural network (a multi-layer perceptron) is trained by conventional stochastic back-propagation algorithm to estimate the probability of 115 context-independent phonetic categories and during recognition it is used as a state output probability estimator. The hybrid HMM/ANN speech recognition system based on Mel Frequency Cepstral Coefficients (MFCC) was developed using CSLU Toolkit. The system was tested on the VDU isolated-word Lithuanian speech corpus and evaluated on a speaker-independent $\sim$750 distinct isolated-word recognition task. The word recognition accuracy obtained was about 86.7%.

Key words: speech recognition, artificial neural networks, hidden Markov models.

1. Introduction

Significant advances have been made in recent years in the area of large vocabulary speaker independent continuous speech recognition. Hidden Markov models (HMMs) are nowadays the most successful modeling approach for speech recognition. A good introduction to HMMs and their use in speech recognition tasks can be found in (Rabiner *et al.*, 1993) and (Jelinek, 1997). In a classical HMM framework (see Fig. 1), observation probabilities are usually represented by mixtures of Gaussians: the observation probability is a weighted sum of normal density functions. Vector quantization can also be used, and discrete probabilities have been estimated. Once the probabilities are obtained, a dynamic programming is performed to find the best path in the HMM, using the Viterbi search algorithm (Rabiner *et al.*, 1993; Jelinek, 1997).

However, standard HMMs suffer from strong assumptions among which are the observation independence assumption stating that the acoustic vectors are not correlated in time and the assumption about underlying HMM state distribution. These assumptions

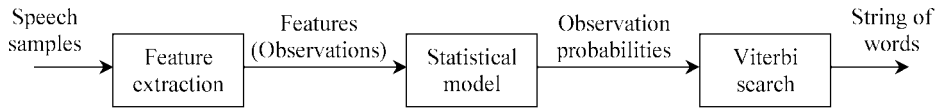

Fig. 1. A simple automatic speech recognition scheme.

can be relaxed by introducing artificial neural networks (ANNs) in the HMM framework[1] (Bourlard *et al.*, 1994), where the neural network (a multi-layer perceptron, or MLP, in our particular case) estimates the state posterior probabilities, which will be used by the HMMs.

The platform for research and development of our recognition system is the CSLU Toolkit[2] (Sutton *et al.*, 1998): a comprehensive software environment that integrates a set of speech-related technologies, including speech recognition, natural language parsing, speech synthesis, and facial animation. The recognition system described in this paper is frame-based hybrid HMM/ANN recognizer with context-independent categories. In this system, frame-based recognition is done with context-independent sub-phonetic states, where the state probability estimation is computed using a neural network.

The hybrid HMM/ANN system has already been successfully applied in connected and continuous English digits recognition tasks (Cosi *et al.*, 1998; Hosom *et al.*, 1998). In this paper, such type of system is tested on the VDU isolated-word Lithuanian speech corpus[3] (Raškinis *et al.*, 2003a), described in more details below. Some good results on this particular corpus have already been published in (Raškinis *et al.*, 2003b; 2003c), where the recognition system consisting of continuous density HMMs and context-dependent models (so-called triphones) was used. We will compare those results with our baseline context-independent hybrid system.

The HMM/ANN technology is unpopular among Lithuanian speech recognizers. The preference being given to dynamic time warping (Lipeika *et al.*, 2002; Tamulevicius *et al.*, 2004), hidden Markov modeling (Raškinis *et al.*, 2003b; 2003c; Laurinciukaite, 2004) or pure ANN (Filipovič, 2003a; 2003b) based speech recognition techniques. As a consequence, the aim of this research was to develop a baseline medium-vocabulary (> 100 words) isolated-word Lithuanian HMM/ANN-based speech recognition system. The research presented in this paper is new and original because it is based on the HMM/ANN recognition paradigm applied to Lithuanian isolated word recognition.

2. Speech Corpus

The experiments described in this paper were based on the VDU isolated-word Lithuanian speech corpus. This corpus contains broadband recordings of 4 speakers (2 males

[1] Known as hybrid HMM/ANN framework.

[2] CSLU Toolkit is developed by the Center for Spoken Language Understanding at Oregon Graduate Institute of Science and Technology, `http://www.cslu.ogi.edu/toolkit`.

[3] VDU isolated-word Lithuanian speech corpus was compiled by the Center of Computational Linguistics at Vytautas Magnus University (VMU), Kaunas, Lithuania.

and 2 females) of Lithuanian, each reading 275 utterances that contain 2–5 phonetically rich words. A close-talk microphone was used for the recordings, under quiet laboratory conditions. The corpus contains 60.6 minutes of speech[4].

The corpus includes time-aligned phone-level and word-level transcriptions as well as a 16-bit, 44.1 kHz speech waveform files for each utterance. (In our experiments speech waveform files were downsampled to 16 kHz). Each recorded waveform is transcribed and verified at the phone-level and word-level by a human labeler. The phone-level transcriptions use the original 116 monophone alphabet that covers a variety of Lithuanian speech sounds (Raškinis, 2001a; 2001b)[5]. The alphabet includes accented and unaccented, short and long vowels (*a*, *à*, *á*, *ã*, *e*, ...), accented and unaccented diphthongs (*ai*, *ái*, *aĩ*, *au*, ...), soft and hard consonants ($b, b', d, d', \ldots$), and accented consonants (*l*, *m*, *n*, *r*, ...)[6]. The word-level transcriptions are also phonetically based and include accent information. For instance, though the word *kasà* meaning *book office* and the word *kãsa* meaning *dig* share the same orthographic form *kasa*, they are considered to be the two distinct words. The corpus is based on a 744-word vocabulary, with almost every word being uttered once by all the speakers.

3. System

The recognition system is based on the baseline CSLU Toolkit frame-based approach illustrated in Fig. 2. The waveform is divided into frames and specific features are computed for each frame. These features describe the spectral envelope of the speech at that frame and at a small number of surrounding frames. This "context window" of features is used to provide the network with information about the dynamics of the speech signal. The features in each frame are classified into phonetic-based categories using a neural network. The outputs of the neural network are used as estimates of the probability, for each phonetic category, that the current frame contains that category. The result of the neural network processing is a $C \times F$ matrix of probabilities, where C is the number of phonetic-based categories, and F is the number of frames in the utterance. The word or words that best match this matrix of probabilities is determined using a Viterbi search, given the vocabulary and grammar constraints.

The system was trained using Mel Frequency Cepstral Coefficients (MFCC) (Davis *et al.*, 1980) and their delta values, with a 16-ms frame size and a 10-ms frame rate. Cepstral-mean subtraction (CMS) (Furui, 1981) was performed, with the mean computed using all frames of data. The input to the network consisted of the features for the frame to be classified, as well as the features for frames at -60, -30, 30, and 60 ms relative to the frame to be classified. Neural network training was done using standard stochastic back-propagation algorithm (Haykin, 1999) on a fully connected feed-forward network (MLP) with number of the input nodes equal to the size of the input vectors, 400 nodes with

[4]The corpus contains 50.9 minutes of pure speech (if pauses between words are not taken into account).
[5]This monophone alphabet is developed at VMU for the purposes of speech recognition research.
[6]Accented consonants are taken from the second half of accented mixed diphthongs (*al*, *am*, *an*, *ar*, ...).

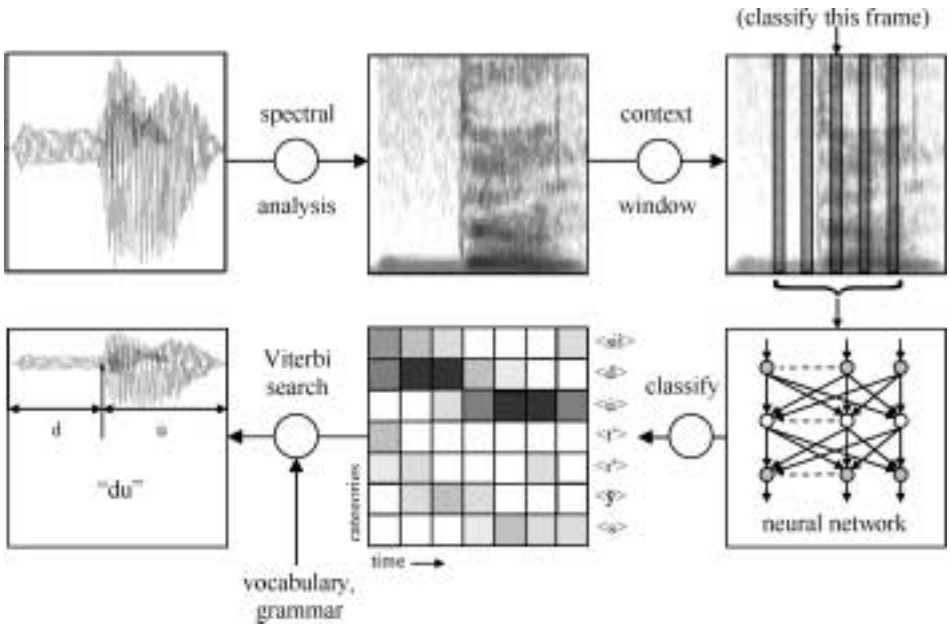

Fig. 2. Overview of baseline CSLU Toolkit frame-based speech recognition using neural networks, illustrating recognition of the word “du” (“two” in Lithuanian).

sigmoidal activations in the single hidden layer, and one node with sigmoidal activation for each context-independent category in the output layer (for a total of 115 output nodes). The weights and biases of the network were initialized randomly. As it is implemented in the CSLU Toolkit, the learning rate η of the network decreased exponentially with each training iteration and the amount of decrease depended on the number of iteration, i.e.,

$$\eta = \frac{\eta_{init}}{1 + 0.2(i-1)},$$

where i is the number of training iteration. The initial learning rate value η_{init} was set to 0.05.

The number of nodes in hidden layer has a strong impact on the performance of an MLP. The more hidden nodes a network has, the more complex decision surfaces it can form, and hence the better classification accuracy it can attain. Beyond a certain number of hidden nodes, however, the network may possess so much modeling power that it can model the idiosyncrasies of the training data if it is trained too long, undermining its performance on testing data.

Depending on the complexity of the problem, the number of nodes in hidden layer varies from 200 for connected digit recognition task (Cosi *et al.*, 1998) to a few thousands for continuous speech recognition task (Renals *et al.*, 1994; Morgan *et al.*, 1995). We evaluated complexity of our problem as a little more complicated than connected digit recognition and decided to use 400 nodes in the hidden layer of the network.

The training was adjusted to use the negative penalty modification proposed by Wei and van Vuuren (Wei *et al.*, 1998). With this method, the non-uniform distribution of context-independent or context-dependent classes that is dependent on the order of words in the training database is compensated by flattening the class priors of infrequently occurring classes; this compensation allows better modeling an utterance in which the order of the words can not be predicted.

During the Viterbi search, transition probabilities were set to be all equally likely, so that no assumptions were made about the *a priori* likelihood of one category following another category. In order to make use of *a priori* information about phonetic durations, and to minimize the insertion of very short words, the search was constrained by specifying minimum duration values for each category, where the minimum value for a category was computed as the value at the second percentile of all duration values. During the search, hypothesized category durations less than the minimum value were penalized by a value proportional to the difference between the minimum duration and the proposed duration. This was done to remove outlier durations that are not representative of their category (such as when phoneme deletion is not accounted for in a word model).

The grammar allowed any number of words in any order, with silence being between words. This grammar allows an unconstrained, free word order and results in a recognition task of high perplexity (geometric mean of the number of words that can follow any other word).

Training was done using 40 iterations, and the "best" network iteration was determined by word-level evaluation of each iteration on the test-set data.

4. Experiments

Two sets of features were evaluated: 13^{th}-order MFCC (including 0^{th} cepstral coefficient) with their delta values (referred to as MFCC13D) and 9^{th}-order MFCC (including 0^{th} cepstral coefficient) with their delta values (MFCC9D). All MFCC features were computed using cepstral-mean subtraction (CMS) processing.

In case of MFCC13D features, for each frame a 130-dimensional vector was constructed. It consists of the 26 features of the frame of interest plus the MFCC13D features of the frames at -60, -30, 30 and 60 ms relative to it. These vectors were the input of the neural network.

Two different cepstral orders (9 and 13) were used to test if the value of 13 is an overrepresentation of the signal. With a sampling rate of 16 kHz, there are on average only 4 formants, and the signal should be adequately represented by 2 cepstral coefficients per formant plus an additional coefficient to approximate the effect of the glottal source. Furthermore, a smaller number of parameters, required by a lower cepstral order, needs to be estimated during training has the potential to make the training data easier to learn.

In order to assess the accuracy of this Lithuanian HMM/ANN speech recognizer several experiments have been conducted:

1. Both MFCC13D and MFCC9D feature sets have been evaluated to see how features affect speech recognizer's performance.

2. Four speaker-independent speech recognition experiments have been performed. Each of these experiments consisted of training speech recognizer on the speech data taken from 3 speakers and testing it on the remaining speech data of 1 speaker who did not take part in the training phase.

5. Results

The performance of speech recognition system can be described in terms of word-level accuracy (WA) defined as:

$$WA = \frac{N - S - I - D}{N} \cdot 100\%,$$

where N is the total number of words in the test-set, S is the number of word substitution errors, I is the number of word insertion errors and D is the number of word deletion errors.

The results of all speech recognition experiments are summarized in the Table 1.

Table 1 reveals encouraging results. The word recognition accuracy (WA) is higher than 84% for all four speaker-independent recognition experiments and achieve on average 86.7%. This WA is achieved regardless of the high perplexity of the speech recognition task. In our experiments, the network training was done using 40 iterations, but it can

Table 1

Speech recognition accuracy using 90-dimensional MFCC9D (above) and 130-dimensional MFCC13D (below) feature vectors. TRN denotes speakers in the training-set, TST denotes speakers in the test-set, ITR denotes the number of the best network training iteration, N denotes the total number of words, S, I, D denotes percentage of word substitutions, insertions and deletions respectively. WA, SA denotes word-level and sentence-level recognition accuracy respectively.

TRN	TST	ITR	N	S	I	D	WA	SA
1, 2, 3	4	21	731	10.40%	0.00%	0.00%	89.60%	76.73%
1, 2, 4	3	26	733	13.51%	0.00%	0.00%	86.49%	68.73%
1, 3, 4	2	22	729	13.31%	0.00%	0.00%	86.69%	71.64%
2, 3, 4	1	13	731	15.87%	0.00%	0.00%	84.13%	66.18%
						Overall:	86.73%	70.82%

TRN	TST	ITR	N	S	I	D	WA	SA
1, 2, 3	4	10	731	17.78%	0.00%	0.00%	82.22%	62.18%
1, 2, 4	3	19	733	18.14%	0.14%	0.14%	81.58%	60.73%
1, 3, 4	2	5	729	15.64%	0.14%	0.00%	84.22%	68.73%
2, 3, 4	1	19	731	16.42%	0.00%	0.00%	83.58%	64.00%
						Overall:	82.90%	63.91%

be seen from the table that, in our case, it would be sufficient to use 30 iterations. It can also be seen that the recognizer trained with the MFCC9D features showed results consistently better than the recognizer trained using the MFCC13D features, with an average 22% reduction in error.

6. Conclusions and Discussion

The development of Lithuanian HMM/ANN speech recognizer for the medium-vocabulary isolated-word speaker-independent speech recognition task has been described in this paper. The hybrid HMM/ANN speech recognition system based on Mel Frequency Cepstral Coefficients (MFCC) was developed using CSLU Toolkit. Speech recognizer was tested experimenting with different feature sets and performing speaker-independent tests. The system was evaluated on the speaker-independent $\sim$750 distinct isolated-word recognition task.

These preliminary experiments evaluated a baseline performance of our hybrid HMM/ANN system tested on the VDU isolated-word Lithuanian speech corpus. The word recognition performance of this system is noticeably better than the results reported in (Raškinis *et al.*, 2003b). The best word recognition accuracy of our system was on average 86.7%. To our knowledge, this is the second best system on that database. The best one is around 91–97% (Raškinis *et al.*, 2003c), where continuous single-Gaussian density HMMs and context-dependent phonetic units (triphones) were used.

Additionally, the use of HMM/ANN system provides some advantages over the continuous Gaussian density HMM system:

- No assumptions need to be made about the independency of the input vector components.
- No assumption is needed about the functional form of the probability density function.
- ANNs are a good match to discriminative objective functions (e.g., mean squared error), and so the probabilities will be optimized to maximize discrimination between sound classes, rather than to most closely match the distributions within each class.

However, using MLPs to estimate posterior probabilities is not as straightforward as it looks. Problems are mainly encountered while training the network:

- Stochastic backpropagation is slow in nature since the training requires much more computation than the more common HMM training paradigm.
- Overtraining is often encountered especially if the number of parameters is large and no cross-validation technique is used.
- Problems of local minima. Since the number of parameters is usually large and since the classes are strongly overlapping, the network can get stuck in local minima.

We are at an early stage of the development of our hybrid HMM/ANN speech recognition system. Much work is still required in order to optimize a number of factors (e.g.,

structure of the network, input window size, network training procedure) that may influence performance of the system. Because of lack of speech data, instead of independent development set, the test set data were used in the designing process of the system. While our results look promising, before drawing any definite conclusions, the experiments and investigations should be extended to much larger speech databases.

Acknowledgments

This research was done using VDU isolated-word Lithuanian speech corpus. The authors would like to thank the Center of Computational Linguistics at Vytautas Magnus University, Arimantas and Gailius Raškinis for permission to use the speech corpus and valuable advices.

References

Bourlard, H., and N. Morgan (1994). *Connectionist Speech Recognition: A Hybrid Approach*. Kluwer Academic Publisher, Boston.

Cosi, P., J.P. Hosom, J. Schalkwyk, S. Sutton and R.A. Cole (1998). Connected digit recognition experiments with the OGI toolkit's neural network and HMM-based recognizers. In *Proceedings of 4th IEEE Workshop on Interactive Voice Technology for Telecommunications Applications*. pp. 135–140.

Davis, S., and P. Mermelstein (1980). Comparison of parametric representations monosyllabic word recognition. *IEEE Transactions on Acoustics, Speech and Signal Processing*, **28**, 357–366.

Filipovič, M. (2003a). Isolated word recognition using neural networks. In *Proceedings of Information Technologies 2003*, Technologija, Kaunas. pp. IX–10-20 (in Lithuanian).

Filipovič, M. (2003b). Research of neural networks based isolated word recognition. *Lithuanian Mathematical Journal*, **43** (spec. issue), 618–624 (in Lithuanian).

Furui, S. (1981). Cepstral analysis techniques for automatic speaker verification. *IEEE Transactions on Acoustics, Speech and Signal Processing*, **29**(2), 254–272.

Haykin, S. (1999). *Neural Networks*: *A Comprehensive Foundation*, 2nd ed. Prentice Hall, New Jersey.

Hosom, J.P., P. Cosi and R.A. Cole (1998). Evaluation and integration of neural-network training techniques for continuous digit recognition. In *Proceedings of International Conference on Spoken Language Processing*, vol. 3. pp. 731–734.

Jelinek, F. (1997). *Statistical Methods for Speech Recognition*. The MIT Press, Cambridge, Massachusetts, London.

Laurinciukaite, S. (2004). On different kinds of speech units based isolated words recognition of Lithuanian language. In *Proceedings of The First Baltic Conference on Human Language Technologies*: *The Baltic Perspective*, Riga. pp. 139–143.

Lipeika, A., J. Lipeikienė and L. Telksnys (2002). Development of isolated word speech recognition system. *Informatica*, **13**(1), 37–46.

Morgan, N., and H. Bourlard (1995). Continuous speech recognition. *IEEE Signal Processing Magazine*, **12**, 24–42.

Rabiner, L.R., and B.-H. Juang (1993). *Fundamentals of Speech Recognition*. Prentice Hall, New Jersey.

Raškinis, A. (2001a). Automatic segmentation and transcription of Lithuanian speech. Research program: *Lithuanian Language in Information Society 2000–2006*, 2001 year report. VDU, Kaunas (in Lithuanian).

Raškinis, A. (2001b). *The Atlas of Phonetic Units*. Manuscript, VDU, Kaunas (in Lithuanian).

Raškinis, A., G. Raškinis and A. Kazlauskienė (2003a). Universal annotated VDU Lithuanian speech corpus. In *Proceedings of Information Technologies 2003*, Technologija, Kaunas. pp. IX–28-34 (in Lithuanian).

Raškinis, G., and D. Raškinienė (2003b). Building medium-vocabulary isolated-word Lithuanian HMM speech recognition system. *Informatica*, **14**(1), 75–84.

Raškinis, G., and D. Raškinienė (2003c). Parameter investigation and optimization for the Lithuanian HMM-based speech recognition system. In *Proceedings of Information Technologies 2003*, Technologija, Kaunas. pp. IX–41-48 (in Lithuanian).

Renals, S., N. Morgan, H. Bourlard, M. Cohen and H. Franco (1994). Connectionist probability estimators in HMM speech recognition. *IEEE Transactions on Speech and Audio Processing*, vol. 2. pp. 161–174.

Sutton, S., R.A. Cole, J. de Villiers, J. Schalkwyk, P. Vermeulen, M. Macon, Y. Yan, E. Kaiser, B. Rundle, K. Shobaki, J.P. Hosom, A. Kain, J. Wouters, D. Massaro and M. Cohen (1998). Universal speech tools: the CSLU toolkit. In *Proceedings of the International Conference on Spoken Language Processing*, vol. 7. pp. 3221–3224.

Tamulevicius, G., and A. Lipeika (2004). Dynamic time warping based speech recognition system. In *Proceedings of The First Baltic Conference on Human Language Technologies*: *The Baltic Perspective*, Riga. pp. 156–161.

Wei, W., and S. Van Vuuren (1998). Improved neural network training of inter-word context units for connected digit recognition. In *Proceedings of International Conference on Acoustics, Speech and Signal Processing*, vol. 1. pp. 497–500.

M. Filipovič received his master of science degree in mathematics from the Vilnius Pedagogical University in 2000. He is a PhD student in Recognition Processes Department at the Institute of Mathematics and Informatics since 2000. His research field is neural network based methods for Lithuanian speech recognition.

A. Lipeika is a doctor of technical sciences, a senior researcher of the Recognition Processes Department at the Institute of Mathematics and Informatics and a professor of the Informatics the Mathematics Department of Vilnius Pedagogical University and an associate professor at Radio-Electronics and Fundamental Sciences Departments of Vilnius Technical University. Scientific interests include processing and recognition of random processes, detection of changes in the properties of random processes, signal processing, speech processing, speech and speaker recognition.

Vidutinio žodyno atskirai tariamų lietuvių šnekos žodžių atpažinimo sistema, grįsta paslėptais Markovo modeliais ir neuroniniais tinklais

Mark FILIPOVIČ, Antanas LIPEIKA

Šiame straipsnyje aprašoma lietuvių šnekos atpažinimo sistema, grindžiama paslėptais Markovo modeliais (PMM) ir dirbtiniais neuroniniais tinklais (DNT). Sistemoje buvo panaudota hibridinė PMM/DNT struktūra, kurioje pilnai sujungtas trijų sluoksnių neuroninis tinklas (daugiasluoksnis perceptronas) buvo mokomas pagal stochastinį atgalinio sklidimo algoritmą įvertinti 115 nuo konteksto nepriklausomų fonetinių kategorijų tikimybes. Atpažinimo metu jis buvo naudojamas PMM būsenų išėjimo tikimybių įvertinimui. Hibridinė PMM/DNT šnekos atpažinimo sistema buvo sukurta naudojant CSLU programinę įrangą. Kaip požymius sistema naudojo melų dažnių skalės kepstro koeficientus (MFCC). Naudojant VDU anotuotą atskirai tariamų lietuvių šnekos žodžių garsyną buvo tiriama požymių skaičiaus įtaka sistemos atpažinimo tikslumui. Atpažinimo sistemos tikslumas buvo vertinamas sprendžiant nuo kalbėtojo nepriklausomą $\sim$750 skirtingų atskirai tariamų žodžių atpažinimo uždavinį. Eksperimentiškai buvo nustatyta, kad sistemos vidutinis žodžių atpažinimo tikslumas yra apie 86,7%.