speech recognition,
language models,
medical information systems

Jerzy SAS[1], Tomasz POREBA[1]

OPTIMAL ACOUSTIC MODEL COMPLEXITY SELECTION IN POLISH MEDICAL SPEECH RECOGNITION

In the paper, the method of acoustic model complexity level selection for automatic speech recognition is proposed. Selection of the appropriate model complexity affects significantly the accuracy of speech recognition. For this reason the selection of the appropriate complexity level is crucial for practical speech recognition applications, where end user effort related to the implementation of speech recognition system is important. We investigated the correlation between speech recognition accuracy and two popular information criteria used in statistical model evaluation: Bayesian Information Criterion and Akaike Information Criterion computed for applied acoustic models. Experiments carried out for language models related to general medicine texts and radiology diagnostic reporting in CT and MR showed strong correlation of speech recognition accuracy and BIC criterion. Using this dependency, the procedure of Gaussian mixture count selection for acoustic model was proposed. Application of this procedure makes it possible to create the acoustic model maximizing the speech recognition accuracy without additional computational costs related to alternative cross-validation approach and without reduction of training set size, which is unavoidable in the case of cross-validation approach.

1. INTRODUCTION

Automatic speech recognition (ASR) in the recent decade became an important tool in medical information systems, both as a method of entering texts into databases and as the utility to control medical devices and software with voice commands. Significant reduction of medical documentation preparation costs as a result of ASR implementation was reported [1,2]. Another advantage of ASR, particularly apparent in diagnostic imaging, is shortening of the report turnaround time, i.e. the time elapsed from the diagnostic examination to signing of the corresponding report by a doctor [3]. On the other hand however, the additional effort form medical staff necessary to begin efficient usage of ASR seems to be a prohibitive factor that restricts the popularity of ASR application in medicine [4].

One of laborious steps that each ASR user is expected to carry out is recording of training utterances that will be used either as a basis for creation of individual speaker acoustic model or as the data for generic acoustic model (AM) adaptation. Our experiences show that in order to create a personalized speaker dependent model, it is necessary to collect at least a few hours of voice recordings [5]. In order to adapt the generic acoustic model, it is sufficient to gather just a few minutes of speaker voice samples. In both cases the speech samples used for AM creation should be used most effectively, so as to maximize the ultimate ASR accuracy. In this way, the end user effort necessary to start using ASR technology will be minimized, while still preserving relatively high recognition accuracy.

Majority of approaches to ASR utilize Hidden Markov Model (HMM), [6-9] as a data structure used by the speech recognizer. Acoustic model in this case defines the probability density functions (pdfs) of observation vector emission in HMM states. The model complexity is determined by the number of distinguishable states and by the number of parameters used in the definition of pdfs. AM training consists in finding such parameter values in the acoustic model, which maximize the likelihood of observation vector sequences observed in recorded training utterances. Typical training method is Baum-Welch procedure [6,7], which approximates maximal likelihood parameter estimation.

The common property of modeling for the sake of pattern recognition is that the model complexity affects the effectiveness of training and in result - the final accuracy of the speech recognizer. When the

[1] Instutute of Informatics, Wroclaw University of Technology, 50-370 Wroclaw, ul. Wyb. Wyspianskiego 27, Poland, email: jerzy.sas@pwr.wroc.pl, tporeba@gmail.com.

model complexity is too low it is not possible to fit the model to actual (unknown) distribution of observations (features). On the other hand, application of the model with high level of complexity usually leads to overfitting to training data. The recognizer recognizes perfectly items from the training set but lacks generalization, i.e. items out of the training set are recognized poorly. In general, the model complexity is defined by the number of parameters which values are adjusted in the training process. In case of HMM applied to ASR, the parameters adjusted in the process of training are mean values and variances of multivariate Gaussian pdfs used to model observation vector distributions in states. In order to model the observation distributions more precisely, Gaussian mixture models (GMMs) are applied in many software packages supporting ASR [8,9]. One of ways to control the acoustic model complexity is to set the number of components in GMMs.

In the works presented in this paper we investigate how the number of Gaussian mixture components in acoustic models for ASR influences the accuracy of the speech recognition. The aim is to elaborate the method of determination of GMM components count which maximizes the accuracy. The trial-on-error method consisting in building models applying various numbers of GMM components and testing their effectiveness is not practically applicable for two reasons:

a) repeating model building and accuracy testing procedure for various model complexities with large training and testing sets is very time-consuming,
b) it requires the subdivision of available speech sample sets into tanning and testing subsets; in result the training set is smaller and speech recognition with obtained model is less accurate than in case when the whole set of speech samples is used for training.

Another possibility is to apply popular information criteria used to evaluate statistical models that are created so as to fit data sets. We tested here two criteria: Bayesian Information Criterion (BIC), [10] and Akaike Information Criterion (AIC), [14]. Experiments carried out and described in the later part of the article show that the optimal number of GMM components can be predicted with BIC criterion. When calculated for the set of acoustic models created with the same set of training utterances, it reaches its minimum for such number of GMM components that is slightly lower than the GMM number for which ASR accuracy is maximized. Based on this observation we proposed the acoustic model training procedure that creates near-optimal model for the most accurate speech recognition.

The article is organized as follows. The next section briefly presents acoustic modeling for ASR and explains the methods used as baselines of concepts presented in the article. Section 3 introduces formulas for AIC and BIC criteria and explains how to interpret it in the context of AM evaluation. In the Section 4 the experiments aimed on testing correlations between investigated information criteria and ASR accuracy in application to medical Polish speech recognition are presented. Finally, the acoustic model training procedure is formulated based on experimental observations and some conclusions are drawn in Section 5.

2. USED METHODS

We consider here the typical approach to ASR based on Hidden Markov Models broadly described in literature [6,7]. The acoustic signal (spoken utterance being recognized) is transformed into a sequence of observation vectors (o_1, o_2, ..., o_t).. Each observation vector consists of 39 elements (MFCC features and their first and second derivatives). The speech process is modeled by a Markov stochastic automaton. The architecture of the automaton consists of three levels. On the first level, individual phonemes (or context-dependent triphones) are modeled by sequences of states. We applied the phoneme model consisting of three emitting states. Entering a state is associated with emission of an observation vector o_i. On the second level, words from the finite dictionary are modeled by concatenation of state sequences corresponding to subsequent phonemes in the phonetic translation of the word. Finally, on the third level, word models are linked into the complete graph, which is a model of the arbitrary sequence of spoken words. The architecture of the complete HMM is presented on Fig. 1. Symbols <*s*> and </*s*> denote here the quasi-words corresponding to the beginning and the end of the utterance correspondingly.

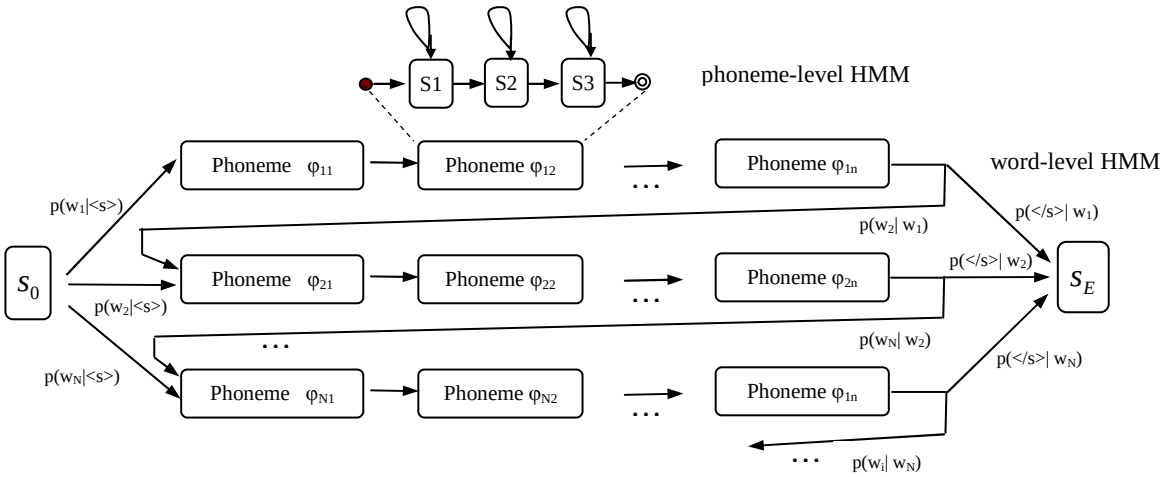

Fig. 1. The architecture of complete HMM for ASR.

The model is stochastic, i.e. transitions between states have associated probabilities. Probabilities of transitions between end states of word models are determined by a *language model*. Probabilities of transitions between subsequent states of phoneme-level model are determined by an *acoustic model*. The acoustic model also determines pdfs of observation emission probabilities $f_i(o)$ for states s_i. In this work we used bigram language model which consists of the set of conditional probabilities $p(w_i|w_j)$. They are the probabilities of occurrences of the word w_i provided that the directly preceding word is w_j. The language model is created by analyzing frequencies of word pairs occurrences in the corpus of domain specific texts.

The utterance recognition consists in finding the most probable state sequence $S^* = (s_O, s_{i_1}^*, s_{i_2}^*, ..., s_{i_t}^*, s_E)$ in the compound HMM conditioned on the sequence of observation vectors (o_1, o_2, ..., o_t) extracted from the acoustic signal of the utterance:

$$S^* = \underset{s_{i_1} s_{i_2} ,...,s_{i_t} \in S^t}{\arg\max} \; P(s_O s_{i_1} s_{i_2} ,...,s_{i_t} s_E \mid o_1, o_2 ,..., o_t), \tag{1}$$

where S^t is the set of state sequences of the length t, such that there is a transition from the state s_i to s_{i+1} in AM for all $i=1,\ldots t-1$. After transforming it to the equivalent form by applying the Bayes formula:

$$S^* = \underset{s_{i_1} s_{i_2} ,...,s_{i_t} \in S^t}{\arg\max} \; P((s_O s_{i_1} s_{i_2} ,...,s_{i_t} s_E) \wedge (o_1, o_2 ,..., o_t)), \tag{2}$$

the optimization problem can be solved using Viterbi dynamic programming algorithm [6] called also Viterbi decoding. Because of the specific structure of HMM model, the found sequence of states which starts in the initial state sO and terminates in the terminal state sE consists of segments corresponding to traversing through individual word-level HMMs. The found sequence of states therefore unambiguously determines the sequence of words which is assumed to be the final result of the utterance recognition.

Probability density functions of observation emission in states are assumed to be Gaussian mixtures, i.e. the pdfs are defined by the formula:

$$b_j(O) = \sum_{m=1}^{M} c_{jm} g(O; \mu_{jm}, U_{jm}), \tag{3}$$

where:

- j - state index,
- O - observation vector random variable,
- M - number of Gaussian components in the mixture,
- c_{jm} - normalization coefficient ($\sum_{m=1}^{M} c_{jm} = 1$ for each state j),
- $g(O;\mu,U)$ - Gaussian pdf with the mean vector μ and the covariance matrix U.

In order to reduce the complexity of the acoustic model, it is assumed that all covariance matrices U_{jm} are diagonal. Typically, all elements of the acoustic model including state transition probabilities and parameters (μ_{jm}, U_{jm}) of pdfs for states for fixed M are estimated using Baum-Welch forward-backward algorithm. Details of the method can be found in [6] and [9].

While the nature of MFCC features in the observation vector o as well as the number of features close to 40 are commonly accepted in ASR systems, the number of components of Gaussian mixtures M is still an open issue. It determines acoustic model complexity, because the number of parameter to be estimated, when training the acoustic model, linearly depends on the number of Gaussian components. Common property of majority of pattern recognition systems is that they are not able to recognize accurately if the model complexity is too low. When the model complexity is too high it tends to overfit data in the training set. In consequence the model lacks the generalization feature. It recognizes objects from the training set accurately, but the recognition performance decreases when objects being recognized are out of the training set. Our experiments with ASR confirmed that this general property also holds in the domain of speech recognition. Therefore the appropriate selection of Gaussian mixture components count is crucial for ASR system performance. Hence, our aim is to elaborate the method which sets the number of GMM components M, so as to maximize the speech recognition accuracy for the given set of traininig utterances.

3. APPLICATION OF AKAIKE AND BAYES INFORMATION CRITERIA TO ASR

Unfortunately, due to the complexity of HMM model, the problem stated above cannot be solved analytically. One reasonable solution seems to be the subdivision of the available set of speech samples (utterances) into training and testing subsets. Then for various values of M, the acoustic model can be created with Baum-Welch procedure using the training subset and the model accuracy can be verified using the testing subset. This approach is however impractical because verification, especially as far as cross-validation approach is used, strongly slows down the process of AM creation. Moreover, the necessity to separate the testing set from available spoken utterances deteriorates the precision of the model and, in consequence, reduces the reliability of the obtained result.

We investigated another approach to fast evaluation of stochastic models based on information criteria. Two most popular criteria have been tested: *Akaike information criterion* (AIC), [10] and *Bayes information criterion* (BIC), [14]. Precise formal derivation of the criteria and the rationale behind its interpretation as stochastic models evaluators can be found in the referenced articles.

AIC measures how good a statistical model fits the training data. Number of parameters which were estimated during model training, and likelihood of the model are necessary for AIC calculation. AIC cannot be used to identify given model as "good" or "bad". It can only compare a set of given models. It is a tool for model selection, not hypothesis verification. The best model, according to this criterion, is the model with lowest AIC rank. According to [14], AIC is defined as follows:

$$AIC = 2k - 2\ln(L_{\max}), \tag{4}$$

where k is the number of parameters estimated in the model and L_{max} is the maximum likelihood value for the ranked model.

Bayesian information criterion (also known as Schwarz criterion - SBIC or SBC) is another model evaluation formula. It is very similar to AIC criterion, although it uses stronger penalty for redundant parameters. The original BIC definition from [14] is:

$$BIC_{org} = \ln(L_{\max}) - 0.5k\ln(n), \tag{5}$$

where k is the number of parameters estimated in the model, n is the size of training sample and L_{max} is the maximum likelihood obtained for the data sample in the trained model. Since BIC is only a relative rank, it could differ by a constant factor. By multiplying (5) by the constant factor *-2* the BIC formula more consistent with (4) can be obtained. For this reason, following the reasoning presented in [15], for the sake of this work BIC criterion is defined as:

$$BIC = k\ln(n) - 2\ln(L_{\max}), \tag{6}$$

The best model, according to BIC criterion, is the model where the rank defined in equation (6) is smallest. AIC and BIC criteria were successfully used in the evaluation of stochastic models in various science domains [11-13] but to our best knowledge there were no such documented attempts to acoustic models evaluation, in particular in applications to Polish speech recognition.

In the application to HMM model evaluation for speech recognition, we evaluate the phoneme state models defined according to (3). For Polish speech we used models corresponding to 40 phonemes in the case of context independent phonemes, what gives 120 emitting states, for which models are created. In the case of context-dependent phonemes (triphones), due to state tying procedure [4], the actual count of states varies depending on the amount of training data. BIC and AIC criteria can be easily computed using formulas (4) and (6) individually for each state of the HMM model. In order to evaluate the complete model, assessments computed for individual states are summed. The number of parameters k in each state is in our approach uniform. The parameters being estimated for each state j and each GMM component m are mean values μ_{jm} and elements of the diagonal of the covariance matrix U_{jm}. Provided that the observation vector consists of 39 features, the number of parameters being estimated for each individual state is:

$$k(M) = 2 * 39 * M\ . \tag{7}$$

The Baum-Welch estimation procedure actually estimates also the near-diagonal elements of state transition matrix for each phoneme or triphone (6 parameters for each phoneme/triphone are used in the assumed architecture of the model) as well as weight factors c_{jm} of GMM components, but they were not taken into account. This is because we focus here rather on state pdfs modeling, which is not directly related to transitions between adjacent states. The number of samples n_j for each state j can be obtained by applying Viterbi alignment to all training utterances and by counting observation vectors assigned to the state. Viterbi alignment consists here in creation of the specific compound HMM for each training utterance so as that only this utterance can be recognized. The utterance-specific HMM is just the concatenation of word models for subsequent words constituting the utterance. The utterance is then recognized with this specific model according to typical recognition rules determined by (1) and (2) with Viterbi decoding. The side-effect of Viterbi decoding algorithm is the segmentation of the recognized observation sequence into the subsequences corresponding to individual states. In this way each observation is assigned to a state and in consequence, for each state j the number of observations n_j assigned to it can be found. Finally, the likelihood L_{max} for the state j can be calculated as:

$$L_{\max}(j) = \prod_{o \in \Theta_j} \sum_{m=1}^{M} c_{jm}\, g(o; \mu_{jm}, U_{jm}), \tag{8}$$

where θ_j is the set of observations assigned to the state j by Viterbi alignment and c_{jm}, μ_{jm} and c_{jm} are parameters determined by the Baum-Welch procedure.

Let J denotes the actual number or distinguishable states in the whole acoustic model Γ. The assessment of the complete acoustic model $\Gamma(M)$ based on M Gaussian components can be obtained by summing assessments related to individual states:

$$\begin{aligned} BIC(\Gamma(M)) &= \sum_{j=1}^{J} (k(M)\ln(n_j) - 2\ln(L_{\max}(j)), \\ AIC(\Gamma(M)) &= 2Jk(M) - 2\sum_{j=1}^{J} \ln(L_{\max}(j)). \end{aligned} \tag{9}$$

Having defined AIC and BIC criteria applied specifically to acoustic models created with the same training set differing in the number of Gaussian mixture component counts, the best model can be

selected by finding such one which minimizes the criterion value. It must be however experimentally verified:

a) whether the model that minimizes the criterion value actually maximizes the ASR accuracy,
b) which of considered criteria (AIC, BIC) correlates better with ASR accuracy.

4. EXPERIMENTS

In order to test the dependence between the accuracy of ASR and the value of AIC and BIC criteria the experiment has been carried out. In the experiment, series of HMM models were created for various numbers of Gaussian mixtures *M*. For each obtained model, the ASR accuracy was tested using the testing set of utterances and information criteria were calculated. The aim is to verify if the maximum of ASR accuracy correlates with the minimum of information criteria, both treated as a function of *M*. HTK package [9] was used in the experiment for model creation and ASR accuracy evaluation. As a speech recognizer we used the Julius decoder [8].

The series of models for increasing *M* were created in an iterative procedure consisting in repeating of the single iteration of Baum-Welch procedure. By the single iteration we mean here the update of model parameters, which is a result of applying Baum-Welch forward-backward procedure for the whole available training utterances set. Baum-Welch procedure just updates the HMM model, so the output of the previous iteration is the input to the next iteration. Initial HMM model used as the input to the first iteration was obtained by estimation the mean vector and diagonal variance matrix with all observations in the training set and by setting these uniform values in all phoneme models. First 9 iterations of Baum-Welch training are executed for Gaussian model with single mixture (*M*=1). Then the number of Gaussian mixtures *M* is increased by 2 and three consecutive iterations of Baum-Welch procedure are executed. This cycle (increase of GMM count by 2 and three iterations of training) is repeated. At the end of each cycle information criteria are calculated and ASR accuracy is tested using a set of verification utterances for the resultant HMM acoustic model.

We performed experiments for 15 speaker-dependent utterances sets being collections of training utterances of the duration ranging from 33 minutes to 4 hours and 59 minutes. Language models were domain-oriented models related to MR and CT diagnostic imaging reports. On the figures below plots of BIC, AIC criteria values and ASR accuracy for selected tests are presented. For remaining test cases the overall relation between ASR accuracy and information criteria values was similar to plots shown in figures 2, 3, 4 and 5.

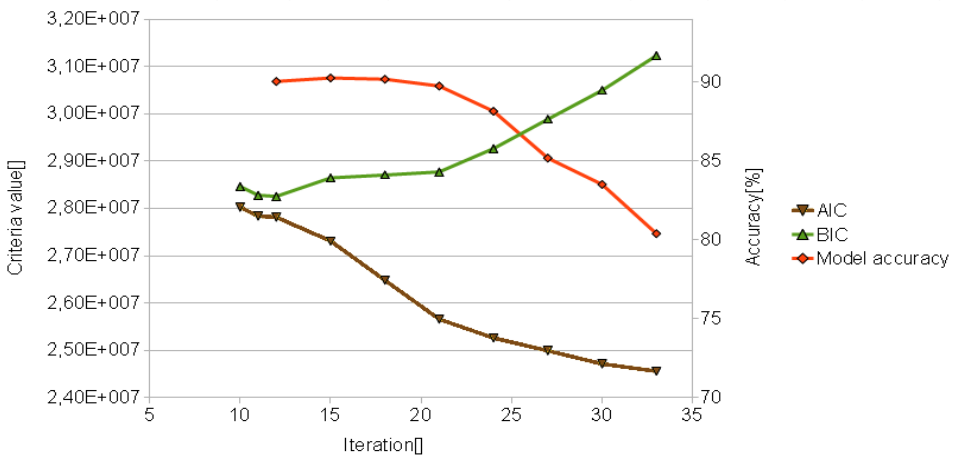

Fig. 2. ASR accuracy vs. AIC and BIC values; 1h 34min of training utterances duration; triphone model.

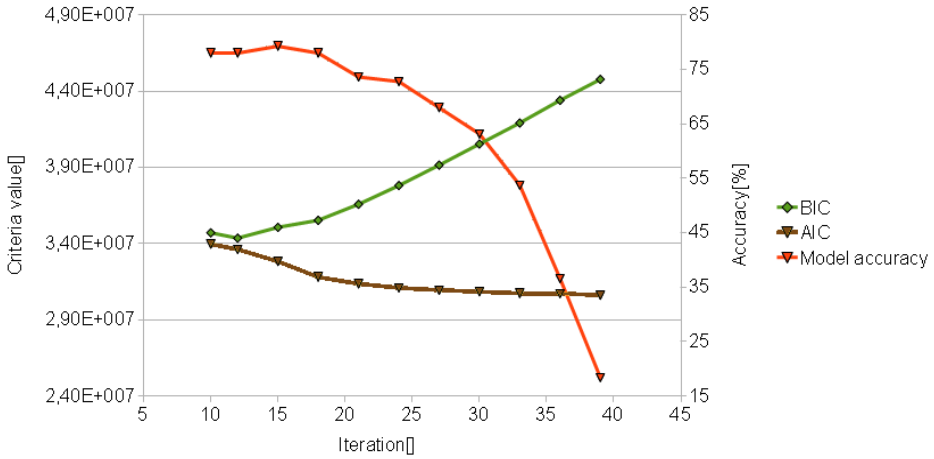

Fig. 3. ASR accuracy vs. AIC and BIC values; 33min of training utterances duration; triphone model.

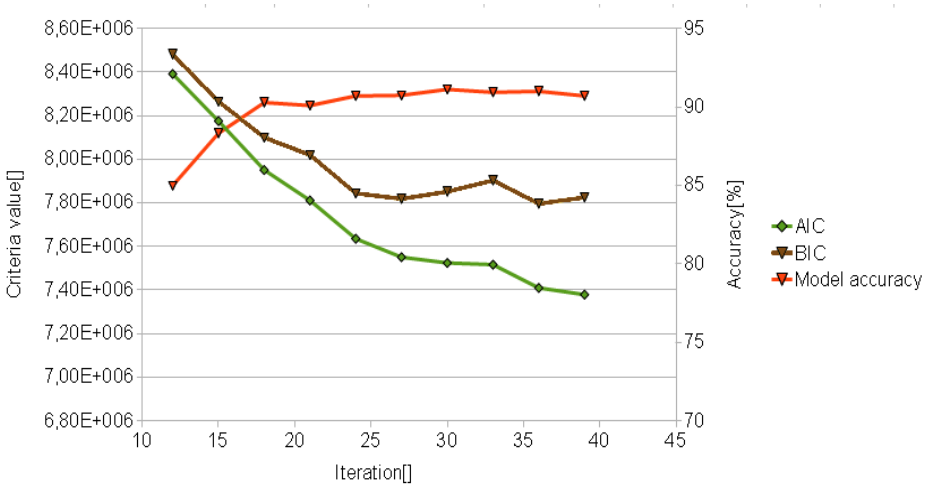

Fig. 4. ASR accuracy vs. AIC and BIC values; 4h 59min of training utterances duration; uniphone model.

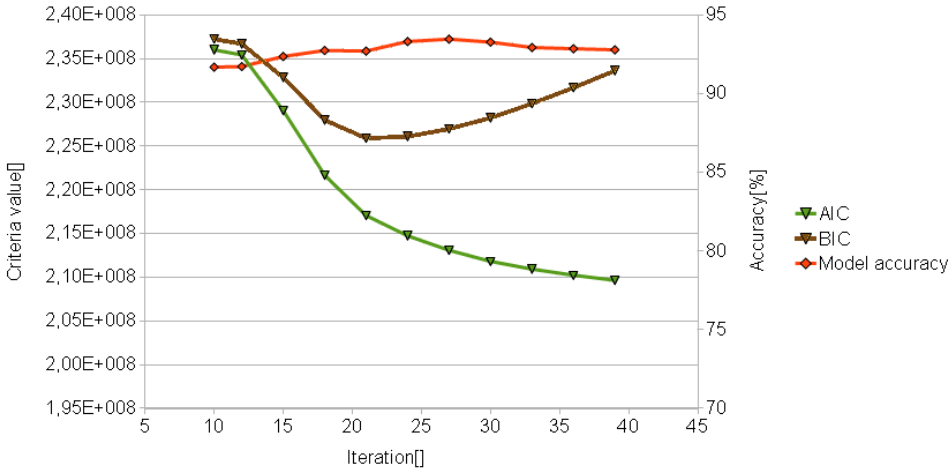

Fig. 5. ASR accuracy vs. AIC and BIC values; 4h 59min of training utterances duration; triphone model.

It can be observed that the minimum of BIC criterion appears in all cases for the model obtained about 6 iterations (2 cycles) before the maximum ASR accuracy is reached. It corresponds to the count of GMM components *M* lower by 4 in relation to *M* value corresponding to the maximal ASR accuracy. This observation makes it possible to formulate practical rule for selection of the near-optimal complexity of the acoustic model: *Continue cyclic increase of GMM counts and re-training while BIC value*

decreases. Then execute two more cycles resulting in further increase of GMM number by 4 and assume the resultant model to be near-optimal.

AIC(M) value considered as a function of *M* does not exhibit properties that make it a predictor of near-optimal complexity of HMM acoustic model. In all cases it constantly decreases with increasing value of *M*. This criterion therefore does not seem to be useful in the problem being considered here.

5. CONCLUSIONS

The aim of works described in this article was to elaborate the method that makes it possible to select near-optimal complexity of acoustic models for ASR. The complexity of the model is determined by the number of parameters being tuned in the process of model creation (training). HMM models used for ASR most commonly define pdfs of observation emision in states as Gaussian mixture models, where the number of components in the mixture can be tuned, so as to maximize ASR accuracy. The method of HMM acoustic model training presented here makes it possible to set near-optimal number of GMM components based on easily-calculable Bayes information criterion. The selection can be achieved without additional computational cost which would be otherwise necessary to carry out the verification procedure based on cross-validation paradigm. The reduction of the training set size, due to the necessity to subdivide the available data into training and testing sets, is not necessary.

Experiments conducted with medical texts corpora proved that Bayes information criterion can be used as relatively accurate predictor of near-optimal model complexity. Although experiments described here were aimed on Polish medical speech recognition applied to diagnostic image reporting, the results probably hold also for other domain-specific language models and for other languages.

BIBLIOGRAPHY

[1] HAO Y., Speech-Recognition Technology in Health Care and Special-Needs Assistance, IEEE Signal Processing Magazine, Vol. 87, 2009.

[2] KOIVIKKO M.P., KAUPINNEN T, AHOVUO J., Improvement of report workflow and productivity using speech recognition - a follow-up study, Journal of Digital Imaging, Vol. 21, No 4, 2008, pp. 378-382.

[3] LANGER S.G., Impact of Speech Recognition on Radiologist Productivity, Journal of Digital Imaging, Vol. 15, No 4, 2002, pp. 203-209.

[4] PEZZULLO J.A., TUNG G.A., ROGG J.M., DAVIS L.M., BRODY J.M., MAYO_SMITH W.W., Voice recognition dictation: radiologist as transcriptionist. Journal of Digital Imaging, Vol. 21, No 4, 2008, pp. 384-389.

[5] HNATKOWSKA B., SAS J., Application of Automatic Speech Recognition to Medical Reports Spoken in Polish, Journal of Medical Informatics & Technologies, Vol 12, 2008, pp. 223-230.

[6] JELINEK F., Statistical Methods for Speech Recognition, MIT Press, Cambridge, Massachusetts, 1997.

[7] JURAFSKY D., MARTIN J., Speech and Language Processing. An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition, Prentice Hall, New Jersey, 2000.

[8] LEE A., KAWAHARA T., SHIKANO K., Julius - an Open Source Real-Time Large Vocabulary Recognition Engine, Proc. of European Conference on Speech Communication and Technology (EUROSPEECH), 2001, pp. 1691-1694.

[9] YOUNG S., EVERMAN G., The HTK Book (for HTK Version 3.4), Cambridge University Engineering Department, 2009.

[10] SCHWARZ G., Estimating the Dimension of a Model, The Annals of Statistics, Vol. 6., No. 2, 1978, pp. 461-464.

[11] LIDDLE A.R., Information Criteria for Astrophysical Model Selection, Monthly Notices of the Royal Astronomical Society: Letters, Vol. 377, No 1, 2007, pp. 74-78.

[12] EVANS J., SULLIVAN J., Approximating Model Probabilities in Bayesian Information Criterion and Decision-Theoretic Approaches to Model Selection in Phylogenetics, Mol. Biol. Evol. Vol. 28, No 1, 2011, pp. 343–349.

[13] ACQUAH H., Comparison of Akaike information criterion (AIC) and Bayesian information criterion (BIC) in selection of an asymmetric price relationship, Journal of Development and Agricultural Economics Vol. 2(1), 2010, pp. 001-006.

[14] AKAIKE H., A new look at the statistical model identification, IEEE Trans. on Automatic Control, , Vol 19, No 6, 1974, pp. 716-723.

[15] BURNHAM K. P., ANDERSON D. R., Multimodel inference: Understanding AIC and BIC in model selection, Sociological Methods and Research, Vol 33, No 2, 2004, pp. 261-304.