PROSODY EVALUATION FOR EMBEDDED SLOVENE SPEECH-SYNTHESIS SYSTEMS

France Mihelič*, Boštjan Vesnicer*, Janez Žibert*, and Elmar Nöth†

*University of Ljubljana, Faculty of Electrical Engineering, Ljubljana, Slovenia
†Universität Erlangen-Nürnberg, Erlangen, Germany

Key words: embedded systems, speech synthesis, HMM acoustic modeling, prosody modeling, speech synthesis evaluation, prosodic tags recognition, support vector machines, RGB kernel

Abstract: This paper describes an evaluation of the prosody modeling in an HMM-based Slovene speech-synthesis system that is suitable for embedded systems due to its relatively small memory footprint. The objective-evaluation procedure is based on the results of the automatic recognition of syntactic-prosodic boundary positions and accented words in the synthetic speech. We have shown that the recognition results represent a close match with the prosodic notations, labeled by the human expert on the natural-speech counterpart produced by the speaker whose speech was used to train the speech-synthesis system. Therefore, the recognition rate of the prosodic events is proposed as an objective evaluation measure for the quality of the prosodic modeling in the speech-synthesis system. The results of the proposed evaluation method are also in accordance with previous subjective-listening evaluation tests, where high scores for the naturalness for such a type of speech synthesis were observed.

Ocenjevenje prozodije za vgrajene sisteme za sintezo slovenskega govora

Kjučne besede: vgrajeni sistemi, sinteza govora, akustično modeliranje s PMM, modeliranje prozodije, evaluacija sinteze govora, razpoznavanje prozodičnih oznak, metoda podpornih vektorjev, jedro RGB

Izvleček: Članek opisuje vrednotenje modeliranja prozodije v sistemu za sintezo slovenskega govora, ki deluje na osnovi prikritih Markovovih modelov (PMM). Zaradi relativno skromne zasedbe pomnilnika programa za sintezo, je tak način zelo uporaben za realizacijo v vgrajenih sistemih. Objektivna metoda vrednotenja je zasnovana na osnovi samodejnega razpoznavanja mest sintaktično-prozodičnih mej in poudarjenih besed v sintetiziranem govoru. Rezultati razpoznavanja kažejo veliko ujemanje med položaji sintaktično-prozodičnih mej in poudarjenimi besedami, ki jih je ekspert označil na pripadajočem naravnem govoru govorca, katerega govor smo uporabili za učenje sistema za sintezo govora. V tem smislu delež pravilno razpoznanih prozodičnih dogodkov predlagamo za merilo uspešnosti prozodičnega oblikovanja v sistemu za sintezo govora. Rezultati predstavljene evaluacije so tudi skladni s predhodnimi subjektivnimi slušnimi ocenjevanji, ki so dala relativno visoke ocene v smislu naravnosti takega načina sinteze govora.

1. Introduction

Modern speech synthesizers are able to achieve high intelligibility; however, they still suffer from rather unnatural speech. Recently, to increase the naturalness of speech-synthesis systems, there has been a noticeable shift from diphone-based toward corpus-based unit-selection speech synthesis /3/. This has been made possible by the rapid increase in the speed and capacity of computer resources. The main idea in unit-selection speech-synthesis systems is to dynamically select appropriate speech units (e.g., diphones) from a large speech database, and in this way reduce the need for signal-manipulation algorithms (e.g., PSOLA) which significantly degrade the quality of the speech.

In such systems the emphasis is more on engineering techniques (searching, optimization, statistical modeling) than on linguistic-rule development /8/. Many of these algorithms are borrowed from the automatic-speech-recognition (ASR) community. For example, hidden Markov models (HMMs) are widely used for the automatic segmentation and labeling of speech databases (e.g. /7/).

However, to achieve a reasonable performance with such corpus-based systems large computational resources are required, and these are often not available when dealing with embedded systems. Since our objective was to develop a speech-synthesis system that should operate with embedded-systems applications, we focused our research on systems with a relatively small memory footprint.

In accordance with current trends an HMM-based approach to speech synthesis for Slovene was implemented. In contrast to other corpus-based speech-synthesis systems, a system using acoustic models trained on a large speech database using HMM techniques provides a relatively small memory footprint, comparable to - or even smaller than - the footprint of embedded diphone-based systems /5, 17/, and demands no special computational load.

Subjective listening-evaluation tests of the HMM-based system showed a surprisingly high level of prosodic quality for such synthesized speech, although no special proso-

dy modeling was used in the system /14/. In our present research we tried to apply the established methods already used in ASR for prosodic-events recognition to evaluate the prosodic content in the synthesized speech.

The rest of the paper is organized as follows. Section 2 overviews the main idea behind the HMM-based speech-synthesis system. In Section 3 we present the prosodic labeling and databases used for the training and the evaluations. Section 4 gives a short description of the prosodic features and classification procedure for the detection of prosodic events in a speech-synthesis system. The results of the experiments are discussed in Section 5. Finally, concluding remarks and plans for future work are presented in the last section.

2 HMM-based speech synthesis

The selected HMM-based approach for speech synthesis differs from other approaches because it uses the statistical framework of HMMs not only for the segmentation and labeling of the database but also as a model for speech production. The method was originally proposed in /11/ and later extended by Yoshimura et al. /15/.

A schematic representation of building and using an HMM-based speech-synthesis system is shown in Figure 1. The top pane shows a *training step* of such a system, where a statistical model of the speech is estimated (middle part). In the bottom pane a *synthesis step* is shown, where the speech signal is generated.

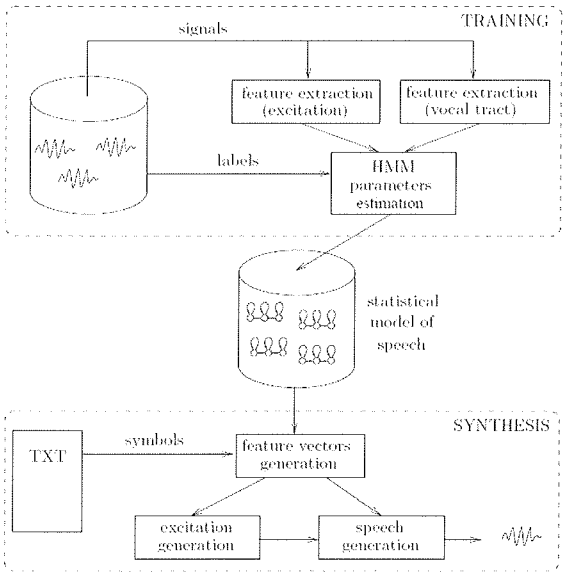

Fig. 1: Schematic diagram of HMM-based training and synthesis procedures.

For a reliable estimation of the parameters of the statistical model a speech parametrization is required. Since we want to be able to synthesize high-quality speech, the parameters should contain enough information for the reconstruction of speech, which should be perceptually similar to the original. For this purpose, the *source-filter* theory of speech production /9/ is applicable. In order to follow this theory, the parameters of the *vocal tract* transfer function and the *excitation* (f_0) need to be estimated. We used MFCC and log f_0 parameters along with their Δ and $\Delta\Delta$ counterparts.

The procedures for estimating the parameters of HMMs are well known from the field of ASR, and can also be applied in our case. A difference here is that additional parameters of the excitation should be modeled. For this purpose, new types of HMMs are introduced. They are called multi-space-distribution (MSD) HMMs, and are presented in /13/.

A detailed description of the speech parametrization and the HMM structure that we used is given in /14/.

For duration modeling a possible solution is to incorporate *explicit* state duration densities (non-parametric or parametric). However, this increases the storage and computational time significantly, and also increases the need for more training material. To avoid those problems we make a simplification and estimate the duration densities only when the training process is already finished /15/. Evaluation tests using this kind of duration prediction give comparable results /14/ to the previously developed two-stage duration model for speech synthesis using the diphone-concatenation technique /4/.

The memory size of the statistical acoustical and duration models, the pronunciation dictionaries and the object code of the current version of the speech-synthesis system was approximately 1.5 MB.

3 Prosody labeling

We have decided to investigate the possibility of generalizing the described speech-synthesis approach to enable some simple prosody modeling. Therefore, the same speech corpora that were used for the training of the HMM acoustic models used in the speech synthesis were manually annotated with syntactic-prosodic boundaries and word accents in the utterances. These corpora consist of 578 sentences (37 minutes of speech) uttered by the speaker 02m from the Slovenian Weather Forecast Speech Database (VNTV) /7/. Recently, when SiBN speech corpora were recorded and annotated /17/ we acquired additional records of weather forecasts from the same speaker consisting of an additional 253 sentences (17 min of speech). These data were also prosodically annotated in the same manner. The syntactic-prosodic boundaries along the lines of /1/ were annotated for the transliterations of speech utterances, and word accents were labeled via acoustic perceptual sessions. We used the same prosodic annotation as in our previously reported work /6/. Similar annotations for Slovene speech were also used by /10/. Three-class annotation was used in each utterance for the prosody boundaries and the word accents. They are listed in Table 1 and Table 2.

Table 1: Syntactic-prosodic boundary labels

M3 :	Clause boundaries
M2 :	Constituent boundaries likely to be marked prosodically
M0 :	Every other word boundary

Table 2: Word accent labels

PA :	The most prominent (primary) accent within a prosodic clause
SA :	All other accented words are marked as carrying a secondary accent
UA	Unaccented words

An example of the labeled text utterance is given below[1]:

V prihodnjih SA *dneh bo vroče* PA M3. *Ob koncu* SA *tedna* PA M2 *pa se bo vročini* SA *pridružila še* M2 *soparnost* SA M3.

English translation: *In the following days it will be hot. During the weekend the hot weather will be accompanied by sultry conditions.*

4 Prosodic features selection and classification

Classification was performed using prosodic feature sets derived from duration segmental characteristics on the word level, speaking rate, energy and pitch. The duration and energy features were additionally normalized by applying normalization procedures based on statistical parameters that were estimated from the training data /2/ (pp. 38-59).

The energy and pitch features were based on the short-term energy and f_0 contour, respectively. Some of the features that were used to describe a pitch contour are shown in Figure 2. Additionally, we used the mean and the median as features /2/ (pp. 59-62).

We derived the same features set (95 features) that were proposed in the experiments on German speech /2/ (page 103). All 95 features were computed on the Slovene speech databases that were used in our evaluations.

The prosody events were detected by support vector machines (SVMs) /19/. In our previous studies a classification with neural networks was also performed /6/, but we gained a significant improvement in terms of recognition scores when SVMs were applied. The LIBSVM software /18/ was used for the training and recognition tests. We used the RBF kernel with $C = 2^3$ and $\gamma = 2^{-5}$. Note that all the data were linearly scaled into the [-1, +1] range during the pre-processing stages.

1 *M0 and UA labels are not indicated in the example.*

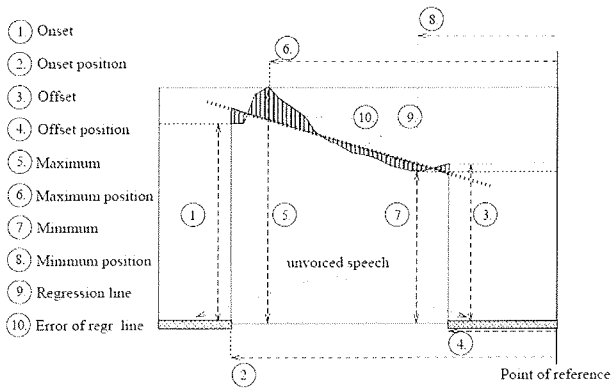

Fig. 2: Example of features used to describe a pitch contour.

5 Experimental results

Our first experiments on the recognition of prosody events were made on the same data set (02m VNTV part) that was used for the training of the HMM-based Slovene speech-synthesis system. We used the data text transcription and the Viterbi forced alignment method to label the speech on the phoneme- and word-duration level. Afterwards, 95 dimensional feature vectors were computed for each word of the uttered speech. In total, 6363 vectors with an accompanying 6363 syntactic-prosodic boundary markers and word-accent markers were determined.

5.1 Cross evaluation on natural speech

The first evaluation was performed on natural speech. The cross evaluation was accomplished by dividing the data into 5 training and test subsets. The cross-evaluation results gave an 81% average overall recognition rate and a 75% class-wise recognition rate for the detection of syntactic-prosodic boundaries, and a 69% overall recognition rate and a 66% class-wise recognition rate for word-accent detection. The results proved the consistency of the speech-data labeling procedure and the appropriateness of using SVM for classification process. The cross-evaluation results from this step also served to determine the applicable pair of RGB-kernel parameters reported in Section 4.

5.2 Recognition of the prosodic events in the synthesized speech counterpart

In our next experiment a new database consisting of the synthesized speech generated from text transcriptions of speech-data recordings used for the training of the speech-synthesis system was acquired. In this way we obtained the synthesized-speech database counterpart of the original natural-speech database (02m VNTV part). We were also able to use the same prosodic markers as in the previous experiment. To determine the prosodic feature vectors for synthesized speech this database was also labeled on the phoneme- and word-duration levels. Surprisingly,

the cross-evaluation check on this database gave entirely comparable results to those obtained from the natural-speech counterpart. And even when we were using the natural-speech database for training and its synthesized counterpart for testing, we still got relatively high recognition scores, with an 83.8% overall recognition rate, a 70.7% class-wise recognition rate for prosodic boundaries, and a 77.5% overall recognition rate and a 69.9% class-wise recognition rate for word accents. The confusion matrices for each class are given in Table 3[2] and Table 4.

Table 3: Confusion matrices for syntactic-prosodic boundaries recognition in absolute and relative figures for synthesized speech. The recognition system was trained on the natural-speech counterpart.

actual/predicted	M0	M2	M3	all actual
M0	3662	429	23	4114
M2	266	1034	43	1343
M3	52	125	151	328
all predicted	3980	1588	217	

actual/predicted	M0	M2	M3
M0	89.0%	10.4%	0.6%
M2	19.8%	77.0%	3.2%
M3	5.9%	38.1%	46.0%

Table 4: Confusion matrices for word-accent recognition in absolute and relative figures for synthesized speech. The recognition system was trained on the natural-speech counterpart.

actual/predicted	UA	SA	PA	all actual
UA	3456	417	151	4024
SA	341	955	137	1433
PA	200	188	518	906
all predicted	3997	1560	806	

actual/predicted	UA	SA	PA
UA	85.9%	10.4%	3.7%
SA	23.8%	66.6%	9.6%
PA	22.1%	20.7%	57.2%

Although, as already mentioned in the Introduction section, subjective listening-evaluation tests of the system showed a high level of prosodic quality /14/, we were surprised with these figures. The relatively high recognition results could be due to the method used to acquire the HMM acoustical models for sub-word units of the speech-synthesis system. In our case, triphone units that represented an acoustic model of each phoneme of Slovene with a specific left and right phoneme context were modeled /14/. Since our training database was very domain specific (nearly consecutive season weather forecasts) consisting of only 770 different words, we could expect that the training material for the specific triphone unit was obtained in a large part from uttered word(s) in a specific prosodic context. Based on this assumption, synthesized words in the same context - we have used text transcriptions of training speech material - possess similar basic prosodic attributes in terms of duration, pitch and energy, and could therefore enable an appropriate prosodic impression.

Since in this case we got a very close match of the prosodic features between natural and synthetic speech, the question arises as to what degradation in terms of automatic prosodic-events recognition could we expect if some different text in the same domain of weather forecasts were to be synthesized?

5.3 Recognition of the prosodic events in synthesized speech from the test set

As mentioned in Section 3, we have recently acquired new speech material from the same speaker that was used for the training of our speech-synthesis system. This speech material includes 267 different words (35%) that were not uttered in the training database, and almost all the uttered sentences - except some short sentences expressing opening and closing remarks - were different from those in the training set. This new database was prosodically annotated. On the basis of its text transcription we were again able to produce its synthesized counterpart, which was additionally labeled on the phoneme and word-duration level. Afterwards, prosodic feature vectors for this new synthetic speech data were computed.

In our next recognition experiments the database that consisted of the synthesized speech from the VNTV database was used to train the SVM classifier, and this new data was used for the recognition tests. In this case we still got a relatively close match between the annotation of the syntactic-prosodic boundaries and the recognition results, which is shown in Table 5[3]. The overall recognition rate was 75% and the class-wise recognition rate was 66%.

2 *In syntactic-prosodic boundaries recognition we did not count the recognition of the M3 marker at the end of each utterance, since the recognition of this marker is trivial.*

3 *As in Table 3, we did not count the recognition of the M3 marker at the end of each utterance.*

Table 5: Confusion matrices for syntactic-prosodic boundaries recognition in absolute and relative figures for the test part of the synthesized speech. The recognition system was trained on the VNTV synthesized speech.

actual/predicted	M0	M2	M3	all actual
M0	1606	303	87	1996
M2	160	328	50	538
M3	31	36	86	153
all predicted	1797	667	223	

actual/predicted	M0	M2	M3
M0	80.5%	15.2%	4.3%
M2	29.7%	61.0%	9.3%
M3	20.3%	23.5%	56.2%

With the recognition of word accents we encountered strong confusion[4] between the primary accent (PA) and other accented words (SA) in comparison with the human labeler's annotations, and therefore a stronger degradation in the recognition scores, where we got a 62% overall recognition rate and 53% class-wise recognition rate. Nevertheless, the automatic distinction between the accented words and those without accent still remains relatively high (Table 6), showing an adequate prosodic content of the test part of the synthesized speech.

Table 6: Confusion matrices for word-accent recognition in absolute and relative figures for the test part of the synthesized speech. The recognition system was trained on the VNTV synthesized speech.

actual/predicted	UA	PA and SA	all actual
UA	1250	417	1783
PA and SA	221	936	1157
all predicted	1471	1353	

actual/predicted	UA	PA and SA
UA	70. 1%	29.9%
PA and SA	19.1%	80.9%

We also separately explored the recognition results for three subsets of test utterances. The first set of utterances (Test 1) consisted of sentences where the speaker used words that were already used in the training process. The second subset was composed of utterances where at least one word was different from the words used in the training stage (Test 2), and in the third subset there were utterances that differed by two or more words (Test 3).

As expected, the recognition results depicted in Tables 7 and 8 depend on the selection of utterances, showing a decreasing recognition rate when more unseen events (words) are included in the test samples. This is also an indication that such a type of evaluation method is reasonably sensitive to small changes in synthesized speech.

Table 7: Recognition results for prosodic-syntactic boundaries in synthesized speech for three different test subsets.

Test sets	Number of utterances	Recognition overall	Recognition class - wise
Test 1	103	76.0%	66.8%
Test 2	150	74.8%	65.4%
Test 3	86	73.8%	62.7%

Table 8: Recognition results of accented words (PA and SA) in synthesized speech for three different test subsets.

Test sets	Number of utterances	Recognition overall	Recognition class - wise
Test 1	103	77.4%	78.5%
Test 2	150	72.9%	74.0%
Test 3	86	72.3%	73.3%

6 Discussion

On the basis of the presented results, the prosodic quality of domain-constrained synthesized speech produced by an HMM-based speech-synthesis system is relatively high, even though no additional prosodic modeling was included in the system. This could be an important issue when dealing with embedded systems where an expansion of the system model usually leads to an additional demand for a larger memory footprint and more computational power.

A further improvement in the prosody modeling could be achieved if the prosodic markers in the training set were to be used to build different prosody-specific acoustic models of sub-word units. Note, however, that in this case the text-to-speech system should be able to extract prosodic markers from the text that is to be synthesized. Some previous studies indicate /1/ that automatic prosodic annotation based on appropriate statistical n-gram modeling gives reasonably good results and could be used for this task. Our evaluations also showed that better results can be expected for limited-domain speech synthesis, which is also the most common synthesis approach when using embedded devices.

4 *This confusion could also partly depend on the in consistency of the human labeler, since there is a 4-year gap between the labeling of the training and test sets.*

After analyzing the recognition results, the automatic recognition of prosodic events was also shown to be an effective tool for the objective evaluation of the performance of speech-synthesis systems on the prosody-modeling level and could therefore also be used for side-by-side comparisons of the different systems. The application of such a type of evaluation in comparison with subjective evaluation tests - where a set of representative listeners should be gathered for each evaluation test, questionnaires prepared, filled in and analyzed - is much less time consuming and could be easily reproduced.

7 Conclusion

We have described experimental results from an automatic recognition of prosodic events in the synthesized speech produced by an HMM-based speech-synthesis system. The results indicate that such kinds of tests could be used as an objective measure for the evaluation of prosody modeling in speech-synthesis systems. They also confirm a relatively good impression of naturalness with HMM-based speech synthesis, which was also noticed during previously performed subjective listening tests. The results of this study also suggest that it could be worthwhile to make use of the presented (or similar) prosodic annotations of training speech corpora for constructing prosody-specific sub-word acoustic models. However, the effectiveness of such an approach is still to be confirmed empirically, which we plan to do in our future experiments.

8 Acknowledgements

The authors wish to thank the Slovenian Ministry of Higher Education, Science, and Technology and the Slovenian Research Agency for co-funding this work under contract no. L2-6277.

9 References

/1/ Batliner A., Kompe R., Kiessling A., Mast M., Niemann H., Nöth E., "M = Syntax + Prosody: A syntactic-prosodic labelling scheme for large spontaneous speech databases", Speech Communication, 25, 1998, pp. 193-222.

/2/ Buckow J., "Multilingual Prosody in Automatic Speech Understanding", Logos Verlag Berlin, 2004.

/3/ Campbell N., Black A., "Prosody and the Selection of Source Units for Concatenative Synthesis", J. van Santen, R. Sproat, J. Olive and J. Hirschberg (Eds.), in Progress in Speech Synthesis, pp. 279-282, Springer Verlag, 1996.

/4/ Gros J., "A two-level duration model for the Slovenian speech", Electrotechnical Review, vol. 66, no. 2, 1999, pp. 92-97.

/5/ Mihelič A., Žganec Gros J., Pavešić N., Žganec M., "Efficient subset selection from phonetically transcribed text corpora for concatenation-based embedded text-to-speech synthesis", Informacije MIDEM 36, nr. 1, 2006, pp. 19-24.

/6/ Mihelič F., Gros J., Nöth E., Warnke V., "Recognition and Labeling of Prosodic Events in Slovenian Speech", Lecture Notes in Artificial Intelligence 1902, Springer, 2000, pp. 165-170.

/7/ Mihelič F., Gros J., Dobrišek S., Žibert J., Pavešić N., "Spoken Language Resources at LUKS of the University of Ljubljana", International Journal of Speech Technology, vol. 6, 2003, pp. 221-232.

/8/ Ostendorf M., Bulyko I., "The Impact of Speech Recognition on Speech Synthesis", Proc. of the IEEE Workshop on Speech Synthesis, 2002.

/9/ Rabiner L., Huang B.-H., "Fundamentals of Speech Recognition", Prentice Hall, Englewood Cliffs, NJ, USA, 1993.

/10/ Stergar J., Horvat B., "Prediction of Symbolic Prosody Breaks with Neural Nets", Informacije MIDEM 32, nr. 3, 2002, pp. 213-218.

/11/ Tokuda K., Kobayashi T., Imai S., "Speech parameter generation from HMM using dynamic features", Proc. of ICASSP, vol. 1, 1995, pp. 660-663.

/12/ Tokuda K., Yoshimura T., Masuko T., Kobayashi T., Kitamura T., "Speech Parameter Generation Algorithms for HMM-based Speech Synthesis", Proc. ICASSP, vol. 3, 2000, pp. 1315-1318.

/13/ Tokuda K., Masuko T., Miyazaki N., Kobayashi T., "Multi-Space Probability Distribution HMM", IEICE Transactions on Information and Systems, vol. E85-D, no. 3, 2002, pp. 455-464.

/14/ Vesnicer B., Mihelič F., "Evaluation of Slovenian HMM-Based Speech Synthesis System", Lecture Notes in Artificial Inteligence 3206, Springer 2004, pp. 513-520.

/15/ Yoshimura T., Tokuda K., Masuko T., Kobayashi T., Kitamura T., "Duration Modeling for HMM-based Speech Synthesis", Proc. ICSLP, vol. 2, 1998, pp. 29-32.

/16/ Zemljak M., Kačič Z., Dobrišek S., Gros J., Weiss P., "Computer-based Symbols for Slovene Speech", Journal for Linguistics and Literary Studies, vol. 2, 2002, pp. 159-294.

/17/ Žganec Gros J., Žganec M., "An Efficient Unit-Selection Method for Embedded Concatenative Speech Synthesis", Informacije MIDEM, vol. 37, 2007, no. 3, pp 158 - 164.

/18/ Žibert J., Mihelič F., "Development of Slovenian broadcast news speech database", Proceedings of Fourth International Conference on Language Resources and Evaluation, Lisbon, Portugal, 2004, pp. 2095-2098.

10 Web References

/19/ Chang C.-C., Lin C.-J., "LIBSVM: a library for support vector machines", http://www.csie.ntu.edu.tw/$\sim$cjlin/libsvm .

/20/ Hsu C.-W., Chang C.-C., Lin C.-J., "A Practical Guide to Support Vector Classification", http://www.csie.ntu.edu.tw/$\sim$cjlin/papers/guide .

prof. dr. France Mihelič,
mag. Boštjan Vesnicer, dr. Janez Žibert
Faculty of Electrical Engineering,
University of Ljubljana
Tržaška 25, 1000 Ljubljana, Slovenija
france.mihelic@fe.uni-lj.si
tel +386 1 4768 841; fax +386 1 4768 316

prof. dr. Elmar Nöth
IMMD5, Universität Erlangen-Nürnberg
Martensstrsse 3, 91058 Erlangen, Germany
noeth@informatik.uni-erlangen.de
tel +49 9131 85 27888; fax +49 9131 303811

Prispelo (Arrived): 20.04.2007 *Sprejeto (Accepted): 15.09.2007*