# GMM Classification of TTS Synthesis: Identification of Original Speaker's Voice*

Jiří Přibil[1][2], Anna Přibilová[3], and Jindřich Matoušek[1]

[1] University of West Bohemia, Faculty of Applied Sciences, Dept. of Cybernetics,
Univerzitní 8, 306 14 Plzeň, Czech Republic
jmatouse@kky.zcu.cz

[2] SAS, Institute of Measurement Science, Dúbravská cesta 9, SK-841 04 Bratislava, Slovakia
Jiri.Pribil@savba.sk

[3] Slovak University of Technology, Faculty of Electrical Engineering & Information Technology, Institute of Electronics and Photonics, Ilkovičova 3, SK-812 19 Bratislava, Slovakia
Anna.Pribilova@stuba.sk

**Abstract.** This paper describes two experiments. The first one deals with evaluation of synthetic speech quality by reverse identification of original speakers whose voices had been used for several Czech text-to-speech (TTS) systems. The second experiment was aimed at evaluation of the influence of voice transformation on the original speaker recognition. The paper further describes an analysis of the influence of initial settings for creation and training of the Gaussian mixture models (GMM), and the influence of different types of used speech features (spectral and/or supra-segmental) on correctness of GMM identification. The stability of the identification process with respect to the duration of the tested sentence (number of the processed frames) was analysed, too.

**Keywords:** quality of synthetic speech, text-to-speech system, GMM classification, statistical analysis

## 1 Introduction

The text-to-speech system (TTS) usually represents the output part of the whole voice communication system with a human-machine interface. The quality, and first of all, the intelligibility of the produced synthetic speech is a basic condition for its usability. Furthermore, it enables setting of a suitable strategy for the dialogue management. Higher quality and naturalness of synthetic speech can be achieved by various methods of speech synthesis, structures of TTS systems, used types of speech inventories, approaches to prosody generation, etc. Several subjective and objective methods are used to verify the quality of produced synthetic speech [1]. The most often used subjective method for giving the feedback information about users' opinion is the listening test. On the other hand, the objective method based on automatic speech

* The work has been supported by the Technology Agency of the Czech Republic, project No. TA01011264, the Grant Agency of the Slovak Academy of Sciences (VEGA 2/0013/14), and the Ministry of Education of the Slovak Republic (KEGA 022STU-4/2014).

P. Sojka et al. (Eds.): TSD 2014, LNAI 8655, pp. 365–373, 2014.
This is preprint prepared by Proceedings editor for Springer International Publishing Switzerland.

recognition system yielding the final evaluation in the form of a recognition score can be used [2]. These recognition systems are often based on neural networks [3], hidden Markov models [4], [5], or Gaussian mixture models (GMM) [6]. The main advantage of these statistical evaluation methods is that they work automatically without human interaction and the obtained results can be numerically judged.

We investigate whether the quality of synthetic speech produced by a TTS system can be evaluated by a reverse identification of the original speaker and whether the re-identification score depends on the used method of speech modelling and synthetic speech production. To verify this hypothesis, the one-level GMM recognizer for identification of the original male and female speakers from the synthetic speech produced by various Czech TTS systems was developed.

Motivation of this work was to analyse further the influence of initial settings in the GMM creation and training phases (number of used mixture components) and different types of used speech features (spectral and/or supra-segmental) on correctness of GMM identification. The GMMs are created and trained on the original speech of the male and female Czech speakers and tested on the speech produced by the Czech TTS systems with several speech synthesis methods. In addition, the stability of the identification process with respect to the duration of the tested sentence (number of the processed frames) is analysed in the paper.

## 2 Method

The Gaussian mixture models can be defined as a linear combination of multiple Gaussian probability distribution functions (GPDFs) of the input data vector [6]

$$f(x) = \sum_{k=1}^{N_{gmix}} \alpha_k P_k(x), \tag{1}$$

where $P_k(x)$ is the GPDF, $\alpha_k$ is a weighting parameter, and $N_{gmix}$ is the number of these functions. For GMM creation, it is necessary to determine the covariance matrix, the vector of mean values, and the weighting parameters from the input training data. Using the expectation-maximization (EM) iteration algorithm, the maximum likelihood function of GMM is found [6]. The performance of the EM algorithm is controlled by the $N_{gmix}$ parameter representing the number of applied mixtures of GPDFs in each of the GMM models. In standard use of the GMM classifier, the resulting score of the model is given by the maximum overall probability for the given class

$$i^* = \arg\max_{1 \le n \le N} score(T, n), \tag{2}$$

where the $score(T, n)$ represents the probability value of the GMM classifier for the models trained for the current $n$-th class in the evaluation process, and $T$ is the input vector of the features obtained from the tested sentence.

For our purpose we need to quantify and compare differences between probability values of the obtained scores; therefore, these values are normalized and the additional parameters are calculated for a subsequent statistical analysis. The next evaluated

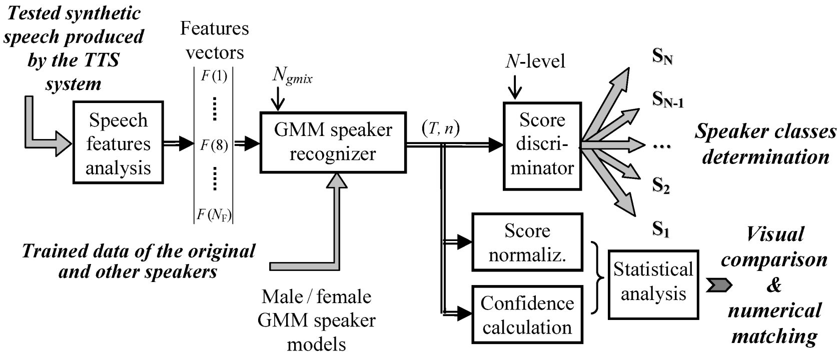
Tested synthetic speech produced by the TTS system
Features vectors
F(1)
F(8)
F(N_F)
Speech features analysis
N_gmix
GMM speaker recognizer
(T, n)
N-level
Score discri-minator
S_N
S_N-1
…
S_2
S_1
Speaker classes determination
Trained data of the original and other speakers
Male / female GMM speaker models
Score normaliz.
Confidence calculation
Statistical analysis
Visual comparison & numerical matching

**Fig. 1.** Block diagram of the GMM recognizer for identification of the original speaker from the synthetic speech produced by the TTS system.

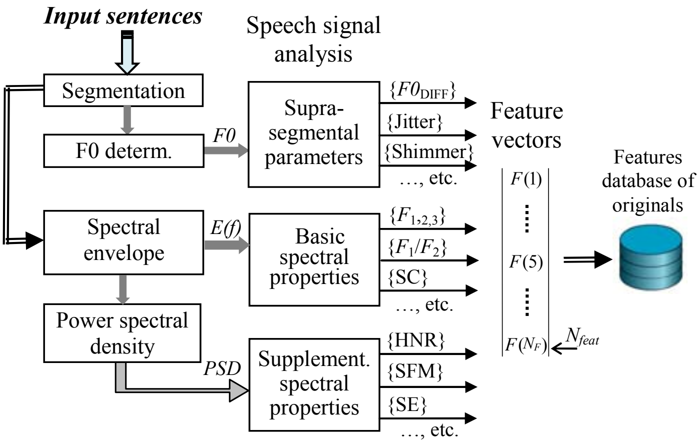

**Fig. 2.** Block diagram of the feature database creation from the spectral properties and supra-segmental parameters of the original speech.

**Table 1.** Basic specification of tested synthetic speech produced by TTS systems.

| Type | Synthesis method | TTS name (specification) | Voice |
|---|---|---|---|
| TTS1$_{M/F}$ | cepstral/harmonic | PC VOX (diphone speech inventory) | Kubec / Ellen |
| TTS2$_{M/F}$ | PSOLA | Epos (triphone speech inventory) | Machac / Violka |
| TTS3$_{M/F}$ | unit selection | ARTIC | Jan / Radka |

parameter is based on the maximum confidence used for selection of features. The confidence measure (CM) gives information how distinctive is the assessment of the given classifier [7]

$$\mathrm{CM} = 1 - \frac{score_{\max 2}}{score_{\max 1}}, \tag{3}$$

where $score_{\max1}$ and $score_{\max2}$ are the highest and the second highest values of the score. The confidence is high when the score for one model is significantly higher than for the other models. On the other hand, the confidence is small when the score is very similar for every model.

One-level GMM recognizer is used for identification of the original speaker from the synthetic speech—see Fig. 1. The precondition of this architecture is the prior correct determination of the gender of the voice (male/female). The speaker recognizer block works with the GMM models that were created and trained using the data of the feature vectors obtained from the speech of the original $N$ speakers. For finding of the optimum recognition accuracy, several values of $N_{gmix}$ are used, the obtained recognition scores are sorted by the absolute size and quantized to $N$ levels corresponding to $N$ output classes in the score discriminator block. In the classification phase, we obtain the scores using the input feature vectors from the tested sentences synthesized by various TTS systems. It means that the highest obtained score represents the synthesized sentences with the values of the speech features that are most similar to those obtained from the original sentences used for GMM training; and the minimum score corresponds to the tested sentence with the greatest differences in comparison to the originals.

The speech signal analysis is performed in the following way: the fundamental frequency F0 is determined from the input sentence after segmentation and weighting. In the next step, the smooth spectral envelope and the power spectral density are computed from the speech frames as shown in the block diagram in Fig. 2. The virtual F0 contour (VF0) is used for determination of the *supra-segmental parameters* describing the microintonation component of speech melody. The differential contour $F0_{\mathrm{DIFF}}$ is obtained by subtraction of mean F0 values and linear trends (including the zero crossings $F0_{\mathrm{ZCR}}$). Further parameters represent microvariations of F0 (jitter) and the variability of the peak-to-peak amplitude (shimmer). The *basic spectral properties* describe the shape of the spectrum obtained from the analysed speech segment. They include the first three formant frequencies and their ratios together with the spectral centroid (SC) and the spectral decrease (tilt). The *supplementary spectral features* are determined from the smoothed magnitude or power spectrum envelope: the spectral flatness measure (SFM), the spectral entropy (SE), and the harmonics-to-noise ratio (HNR) providing an indication of the overall periodicity of the speech signal. Obtained values in the form of the feature vectors with the length $N_{feat}$ are subsequently stored in a database containing the features of the original speakers in dependence on the voice type (male/female) for further processing.

## 3 Material, Experiments, and Results

The speech material for GMM creation and training consists of the short sentences with duration from 0.5 to 2.5 seconds, resampled at 16 kHz, representing the original speech in Czech language uttered by five male and five female speakers (already used in another research [8])—typically 50 sentences per speaker. We have also additional sentences originated from the speaker whose voice was used for building of the speech corpus (male/female) of the tested TTS systems with basic parameters given in Table 1. So we finally have the speech material consisting of 6+6 speakers for testing in each of our identification experiments (designated as Orig1-$6_{\mathrm{M/F}}$) where the voice number indicates the original speech material for the tested TTS system ($\mathrm{Orig1_M}$ is the source speech material for synthesis of the voice $\mathrm{TTS1_M}$ and so on). As regards the synthetic speech (TTS1-$3_{\mathrm{M/F}}$), the database consists of testing sets including 25 short sentences produced by the TTS systems using different types of speech modelling (cepstral [9], harmonic [10], PSOLA [11], unit selection [12], [13]).

The main experiment was focused on identification of the original speaker from the synthetic speech produced by the Czech TTS systems. The second experiment was aimed at evaluation of the influence of voice transformation on the original speaker recognition. The speech material used here was the synthetic speech produced by the TTS system PCVOX—implemented in the special aids for blind and partially sighted people [14], [15]. Four synthetic voices were compared: the basic male voice (synthesis from the original speaker $\mathrm{TTS1_M}$—see Table 1) and the transformed voices of a young male (Tr-young), a female (Tr-female) and a child (Tr-child) [8]. In addition, our research was aimed at investigation of:

- influence of the number of used mixtures (from 2 to 8) on GMM evaluation,
- influence of the used feature set (P1-P3),
- stability of the identification process depending on the tested sentence duration.

The input data vector for GMM training and classification contains the supra-segmental parameters {VF0, $\mathrm{F0_{DIFF}}$, $\mathrm{F0_{ZCR}}$, jitter, and shimmer}, the basic spectral features determined from the spectral envelopes $\{F_{1,2,3}$, $F_1/F_2$, SC, and tilt$\}$, and the supplementary spectral parameters {HNR, SFM, SE}. In the case of the spectral features, the basic statistical parameters—mean values and standard deviations (std)—were used as the representative values in the feature vectors for GMM evaluation. For implementation of the supra-segmental parameters of speech, the statistical types—median values, range of values, std, and/or relative maximum and minimum were used in the feature vectors. The length of the input feature vector $N_{feat} = 16$ was experimentally chosen in correspondence with the obtained results of our previous research [16]. The three tested feature sets were: P1 consisting of the basic spectral features together with the supra-segmental parameters, P2 consisting of the supplementary spectral features and the supra-segmental parameters, and P3 being a mix of the basic and the supplementary spectral features with the prosodic parameters.

As regards the GMM classifier, the simple diagonal covariance matrix of mixture models was applied in this identification experiment. The basic functions from the Ian T. Nabney "Netlab" pattern analysis toolbox [17] were used for creation of the GMM models, data training, and classification.

The obtained results of the GMM identification are presented in a graphical form (for visual comparison) and also as the values for numerical matching separately with respect to the TTS voice gender. The used order of tables and figures corresponds to the course and evaluation of the performed experiments. If not otherwise stated, the presented graphs and tables were determined with the following parameter setting: $N_{gmix} = 5$, feature set P2.

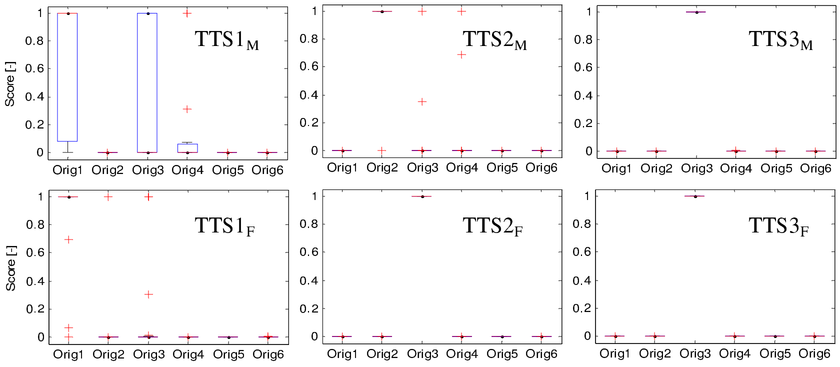

**Fig. 3.** The boxplot of the basic statistical parameters of the normalised GMM score: for male (upper set) and female (bottom set) TTS voices.

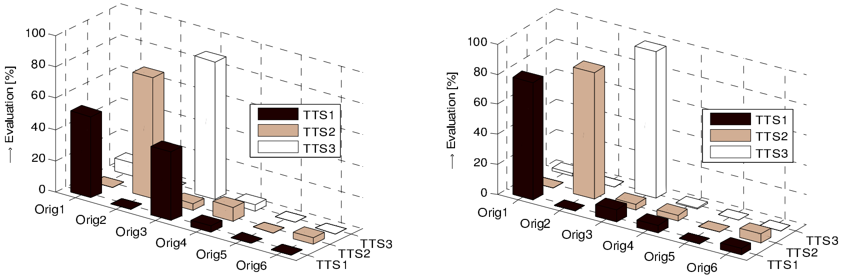

**Fig. 4.** Confusion matrices of original speaker identification for male (left) and female (right) TTS voices of six originals, three TTS synthesis systems.

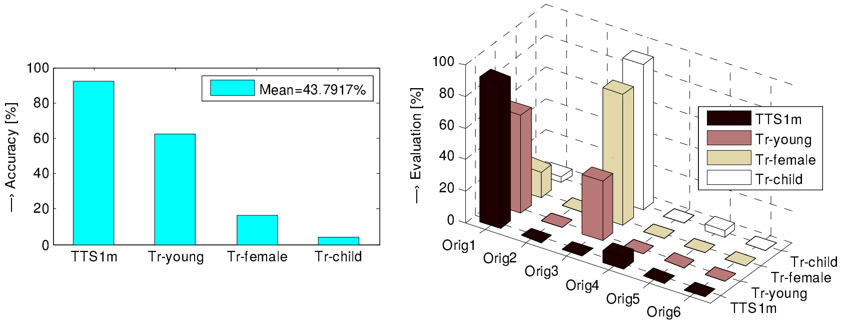

**Fig. 5.** Results of the second identification experiment of TTS synthesis with the male voice and its conversion to young male, female, and childish voice: a bar graph of the mean recognition accuracy (left), a detailed confusion matrix (right).

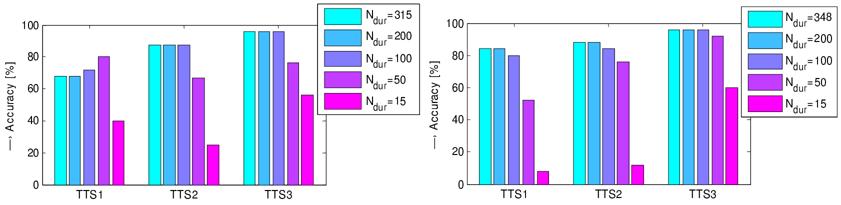

**Fig. 6.** Influence of the tested sentence lengths Ndur in [frames] on the original speaker GMM identification accuracy: for male (left) and female (right) TTS voices.

## 4 Discussion and Conclusion

The performed experiments have shown that there exists a principal influence of a chosen type of parameters in a feature vector on the stability and accuracy of the GMM identification. The best results are produced by the feature set P3 consisting of a mix of spectral and prosodic features, the worst results correspond to the set P2 (see results in Table 4) when only the supplementary spectral and supra-segmental features were used. Therefore, the detailed comparison for this worst case was performed next. For the original speaker with the worst identification from the synthetic speech the confidence measure has the lowest value of the minimum and the highest value of the standard deviation—see boxplots of the basic statistical parameters of the normalised GMM score in Fig. 3 and CM values in Table 2. In the case of the TTS system PCVOX producing voices with worse quality and naturalness, also the lowest original speaker identification accuracy was achieved (49% for male and 77% for male). Relatively great differences between male and female synthesis were probably caused by different used speech models (cepstral one in the case of the male voice, and the harmonic one for the female voice). The last tested TTS system using the unit selection synthesis method (TTS3$_{M/F}$) has the quality of the synthetic speech very near the original voice as shown by the best results of identification accuracy (96% for male and 95% for female voices)—see the corresponding 2D representation of confusion matrices in Fig. 4. From the second recognition experiment follows that the obtained score values correspond to the degree of the voices transformation: the highest score corresponds to the basic male voice and

**Table 2.** Basic statistical parameters of the CM values calculated from the GMM score for the male and female TTS voices.

| TTS voice | Min$^{*)}$ | Mean | Std |
|---|---|---|---|
| TTS1$_{M/F}$ | 0.311 / 0.807 | 0.937 / 0.948 | 0.1522 / 0.0455 |
| TTS2$_{M/F}$ | 0.647 / 0.943 | 0.958 / 0.996 | 0.0985 / 0.0127 |
| TTS3$_{M/F}$ | 0.996 / 0.980 | 0.999 / 0.999 | 0.0040 / 0.0007 |

$^{*)}$ Maximum is equal to 1 in all cases.

**Table 3.** Mean original speaker GMM identification accuracy in [%] in dependence on the number of used mixtures.

| TTS voice | $N_{gmix}$ =2 | $N_{gmix}$ =3 | $N_{gmix}$ =4 | $N_{gmix}$ =5 | $N_{gmix}$ =6 | $N_{gmix}$ =7 | $N_{gmix}$ =8 |
|---|---|---|---|---|---|---|---|
| TTS1$_{M/F}$ | 42/70 | 42/72 | 50/76 | 52/78 | 54/80 | 50/82 | 54/80 |
| TTS2$_{M/F}$ | 72/84 | 72/86 | 72/88 | 76/84 | 72/86 | 75/92 | 72/89 |
| TTS3$_{M/F}$ | 87/96 | 85/96 | 88 /97 | 88/98 | 88/99 | 89/99 | 88/99 |

**Table 4.** Summary results of original speaker GMM identification accuracy in [%] for different types of used feature vectors.

| TTS voice/ feature vector type | P1 | P2 | P3 |
|---|---|---|---|
| TTS1$_{M/F}$ | 48.0/76.3 | 52.0/78.2 | 68.0/82.5 |
| TTS2$_{M/F}$ | 75.0/84.0 | 75.5/84.1 | 87.5/95.5 |
| TTS3$_{M/F}$ | 100/92.8 | 87.5/98.2 | 100/100 |

the lowest one to the transformation to the childish voice (see the bar-graph and the 2D confusion matrix in Fig. 5). Contrary to our expectations, the number of used mixtures has not great significance (see values in Table 3), so the setting of $N_{gmix} = 5$ was chosen for next processing and comparison. Finally, it can be said that the results obtained in this way are in good correspondence with the predicted working hypothesis. The last part of our experiment showed that the length limitation of the processed speech signal practically does not play essential role (see Fig. 6) because our GMM original speaker identifier was developed for testing of continuous speech (i.e. sentences—not isolated words).

Increase of the original speaker identification accuracy can be expected if the full covariance matrix is used for GMM model creation, training, and employment in the classification process, so in near future we will compare approaches using the diagonal and the full covariance matrices.

## References

1. Blauert, J., Jekosch, U.: A Layer Model of Sound Quality. Journal of the Audio Engineering Society 60, 4–12 (2012)
2. Kondo, K.: Subjective Quality Measurement of Speech: Its Evaluation, Estimation and Applications. Springer (2012)
3. Zelinka, J., Trmal, J., Müller, L.: On Context-Dependent Neural Networks and Speaker Adaptation. Proc. IEEE Conf. Signal Processing 2012, Beijing, China, pp. 515–518 (2012)
4. Pražák, A., Psutka, J.V., Psutka, J., Loose, Z.: Towards Live Subtitling of TV Ice-Hockey Commentary. Proc. SIGMAP 2013, Reykjavík, Iceland, pp. 151–155 (2013)
5. Jeong, Y.: Joint Speaker and Environment Adaptation Using TensorVoice for Robust Speech Recognition. Speech Communication 58, 1–10 (2014)
6. Reynolds, D.A., Rose, R.C.: Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models. IEEE Trans. on Speech and Audio Processing 3, 72–83 (1995)
7. Vondra, M., Vích, R.: Evaluation of Speech Emotion Classification Based on GMM and Data Fusion. Esposito A. and Vích R. (Eds.) Cross-Modal Analysis, LNAI 5641, 98–105, Springer (2009)
8. Přibilová, A., Přibil, J.: Non-Linear Frequency Scale Mapping for Voice Conversion in Text-to-Speech System with Cepstral Description. Speech Commun. 48(12), 1691–1703 (2006)
9. Vích, R., Přibil, J., Smékal, Z.: New Cepstral Zero-Pole Vocal Tract Models for TTS Synthesis. Proc. IEEE Region 8 EUROCON 2001, vol. 2, pp. 458–462 (2001)
10. Přibilová, A., Přibil, J.: Harmonic Model for Female Voice Emotional Synthesis. In: Fierrez, J. et al. (Eds.) Biometric ID Management and Multimodal Communication, LNCS 5707, 41–48, Springer (2009)
11. Horák, P.: Czech Pitch Contour Modeling Using Linear Prediction. In: Sojka, P. et al. (Eds.) TSD 2008, LNCS 5246, 333–339, Springer (2008)
12. Tihelka, D., Kala, J., Matoušek, J.: Enhancements of Viterbi Search for Fast Unit Selection Synthesis. Proc. INTERSPEECH 2010, Makuhari, Japan, pp. 174–177 (2010)
13. Romport, J., Matoušek, J.: Formal Prosodic Structures and Their Application in NLP. In: Matoušek V. et al. (Eds.) TSD 2005. LNCS 3658, ipp. 371–378, Springer (2005)
14. Přibil, J., Přibilová, A.: Czech TTS Engine for BraillePen Device Based on Pocket PC Platform. In: Proc. Conf. Electronic Speech Signal Processing (ESSP 2005), pp. 402–408 (2005)

15. Personal Computer Voices: PCVOX. Spektra v.d.n., accessed 5 February 2014 at `http://www.pcvox.cz/pcvox/pcvox-index.html`
16. Přibil, J., Přibilová, A.: Evaluation of Influence of Spectral and Prosodic Features on GMM Classification of Czech and Slovak Emotional Speech. EURASIP Journal on Audio, Speech, and Music Processing 2013(8), 1–22 (2013)
17. Nabney, I.T.: Netlab Pattern Analysis Toolbox. Retrieved 2 October 2013, from `http://www.mathworks.com/matlabcentral/fileexchange/2654-netlab`