ROBUST NEURAL NETWORK-BASED ESTIMATION OF ARTICULATORY FEATURES FOR CZECH

*Petr Mizera, Petr Pollak**

Abstract: The article describes a neural network-based articulatory feature (AF) estimation for the Czech speech. First, the relationship between AFs and a Czech phone inventory is defined, and then the estimation based on the MLP neural networks is done. The usage of several speech representations on the input of the MLP classifiers is proposed with the purpose to obtain a robust AF estimation. The realized experiments have proved that an ANN- based AF estimation works very reliably especially in a low noise environment. Moreover, in case the number of neurons in a hidden layer is increased and if the temporal context DCT-TRAP features are used on the input of the MLP network, the AF classification works accurately also for the signals collected in the environments with a high background noise.

Key words: *Speech recognition, articulatory features, robust estimation, neural networks, MLP, temporal patterns, TRAP*

Received: December 6, 2013 **DOI:** 10.14311/NNW.2014.24.027
Revised and accepted: October 3, 2014

1. Introduction

Speech technology applications are nowadays used in many situations, when a natural voice input is used in communication between humans and machines or in the detection of various phenomena using automated speech analysis. As in other research fields, various subparts of speech technology systems use artificial neural networks for the classification purposes [3, 9, 25]. ANNs can be found, for example, as phone classifiers in the TempoRAl Patterns (TRAP) speech feature extraction [6, 8], as voice activity detectors, as a subpart of the combined Artificial Neural Network and Hidden Markov Model (ANN/HMM) classifiers, in language modelling for continuous speech recognition, or as Kohonen self-organized maps used for selected classification purposes [33], etc.

To improve the robustness of spontaneous or noisy speech recognition, Articulatory Features (AFs) were used in various tasks in continuous speech recognition.

*Petr Mizera, Petr Pollak – Corresponding author, Czech Technical University in Prague, Faculty of Electrical Engineering, CTU FEE K13131, Technicka 2, 166 27 Praha 6, Czech Republic

The most important publications describe the basic definition of AFs [1, 16], applications in continuous speech recognition based on combining standard acoustic and articulatory features [19, 22, 24, 28], and finally also their estimation for which the neural network-based classifiers are used most typically [14]. An automatic estimation of AFs can be used also in the basic phonetic research.

This paper describes the first study of the neural network-based AF estimation for the Czech language utterances together with the study of the achieved accuracy in various speech representations used for the AF classification that can be applied generally for any language. It was proved by the previous research that AFs are generally language-independent, yet particular features are defined for different phone inventories in various languages. Moreover, the extensive study [24] describing the cross-lingual AF detection for English, German, Spanish, Japanese, and Chinese showed, that the best results were always obtained with the AF detectors matched to particular language. Expecting an AF application during the recognition and the phonetic segmentation focused on the Czech spontaneous speech, one purpose of this work is to define the AFs for the Czech phone inventory, because most previously published works describe the research done for English [2, 14, 18] (sometimes also for other languages, but not yet for Czech). The other important part of this work concerns the optimization of an ANN size, thus a detailed analysis of the achieved accuracy of the AF estimation depending on an ANN size is presented here.

Finally, further application of AFs can be expected when the analyzed speech contains a strong background noise, so another purpose of this paper is to provide the analysis of an AF estimation accuracy under the clean vs. noisy conditions. Most works do not deal with the data gathered under adverse background conditions; the experiments in published works are usually conducted with the TIMIT database which contains speech data recorded under the low noise conditions [13, 14]. Some analysis of the noise robustness can be found in [17], where the experiments are also performed with the noisy data from the Verbmobil database using special MODSPEC preprocessing [15] of the input features.

2. Articulatory features

AFs can be represented by some phonological feature sets such as binary features, multi-valued features, and articulatory gestures etc., see [14, 20]. In this paper we use the term articulatory features to refer to the multi-valued features representation using phonetic categories. In comparison to the standard acoustic features, e.g. Mel-Frequency Cepstral Coefficients (MFCC) or Perceptual Linear Prediction (PLP) [10] which provide the information about spectral characteristics of the uttered speech, the AFs imply important information related to the speech production, mainly to the voicing, manner of articulation, place of articulation, etc. Since this information is generally suprasegmental, i.e. AFs are the same across a group of phones, it was proved that it can help in adverse conditions when phone-based models fail due to more casual speech or present a background noise. It is described in details in [21], which deals with manual transcription of the speech at the AF level and where the potential of AF for better modelling of co-articulation in a conversational speech is also discussed. In [32] the authors

			manner											
			stop			*affricates*		*fricatives*					*approximants*	
			plosives		*nasals*					*trills*			*lateral*	*glides*
place	*labial*	*bilabial*	*p*	b	m									
		labiodental			*F*			f	v					
	alveolar	*prealveolar*	*t*	d	n	t_s	d_z	s	z	Q\	P\	r	l	
		postalveolar				*t_S*	d_Z	S	Z					
	palatal		*c*	J	J\									j
	velar		*k*	g	N			x						
	glottal								*h*					
sonority (sonors/noises)			No	No	So	No	No	No	No	No	No	So	So	So
voicing (voiced/unvoiced)			U	V	V	U	V	N	V	U	V	V	V	V

Tab. I *The phonetic categorization of the Czech consonants.*

		manner		
		front	*central*	*back*
place	*close*	*i i:*		u u:
	mid	*e e:*	@	o o:
	open		*a a:*	
rounding		*unrounded*		*rounded*

Tab. II *The phonetic categorization of Czech vowels.*

proved that by incorporating acoustic-phonetic information, the automatic speech recognition performance improved; in this study the manner and place of articulation were considered. For the conversational and hyper-articulated speech, the robustness of AFs against the style of speech was proved in [24].

2.1 Articulatory features for Czech

As any similar study of AFs for Czech has not been published yet, their specific language definition is presented in the following paragraphs. A standard inventory of the phones for Czech defined by the SAMPA standard [36] consists of 49 phones including several rare allophones together with the schwa and glottal stop which are not in Czech canonical pronunciation. Within this introductory work with AFs for Czech we use the same set of phones as it is standardized and used for the Czech speech recognition systems. This set does not contain syllabic variants of consonants, i.e. phones "*m=*, *l=*, *r=*", voiced phone "*G*" which appears only in very special contexts at word boundaries, and the glottal stop "*?*" which does not regularly appear in the Czech pronunciation.

The resulting phone inventory consists of 44 phones which can be categorized into phonetic classes according to the methodology described in [2, 13, 18] for English together with the application of standard conventions for Czech defined by [26, 35]. More particular details of the Czech vowel and consonant categorization is described in Tabs. I and II and a complete overview of the AFs used for the complete phone inventory of Czech is presented in Tab. III.

Phones	*Voicing*	*Place_con*	*Place_vow*	*Manner_con*	*Manner_vow*	*Round*	*Sonor*
i	+	nil	front	nil	high	−	nil
e	+	nil	front	nil	middle	−	nil
a	+	nil	central	nil	low	−	nil
o	+	nil	back	nil	middle	+	nil
u	+	nil	back	nil	high	+	nil
i:	+	nil	front	nil	high	−	nil
e:	+	nil	front	nil	middle	−	nil
a:	+	nil	central	nil	low	−	nil
o:	+	nil	back	nil	middle	+	nil
u:	+	nil	back	nil	high	+	nil
o_u	+	nil	back	nil	middle	+	nil
a_u	+	nil	central	nil	low	−	nil
e_u	+	nil	front	nil	middle	−	nil
@	+	nil	central	nil	middle	nil	nil
p	−	bilabial	nil	stop	nil	nil	−
b	+	bilabial	nil	stop	nil	nil	−
t	−	prealveolar	nil	stop	nil	nil	−
d	+	prealveolar	nil	stop	nil	nil	−
c	−	palatal	nil	stop	nil	nil	−
J	+	palatal	nil	stop	nil	nil	−
k	−	velar	nil	stop	nil	nil	−
g	+	velar	nil	stop	nil	nil	−
t_s	−	prealveolar	nil	affricates	nil	nil	−
d_z	+	prealveolar	nil	affricates	nil	nil	−
t_S	−	postalveolar	nil	affricates	nil	nil	−
d_Z	+	postalveolar	nil	affricates	nil	nil	−
f	−	labiodental	nil	fricatives	nil	nil	−
v	+	labiodental	nil	fricatives	nil	nil	−
s	−	prealveolar	nil	fricatives	nil	nil	−
z	+	prealveolar	nil	fricatives	nil	nil	−
Q	−	prealveolar	nil	trills	nil	nil	−
P	+	prealveolar	nil	trills	nil	nil	+
S	−	postalveolar	nil	fricatives	nil	nil	−
Z	+	postalveolar	nil	fricatives	nil	nil	−
j	+	palatal	nil	glides	nil	nil	+
x	−	velar	nil	fricatives	nil	nil	−
h	−	glottal	nil	fricatives	nil	nil	−
r	+	prealveolar	nil	trills	nil	nil	+
l	+	prealveolar	nil	lateral	nil	nil	+
m	+	bilabial	nil	nasals	nil	nil	+
n	+	prealveolar	nil	nasals	nil	nil	+
N	+	velar	nil	nasals	nil	nil	+
J	+	palatal	nil	nasals	nil	nil	+
F	+	labiodental	nil	nasals	nil	nil	+

Tab. III *Summary of the articulatory features of particular Czech phones.*

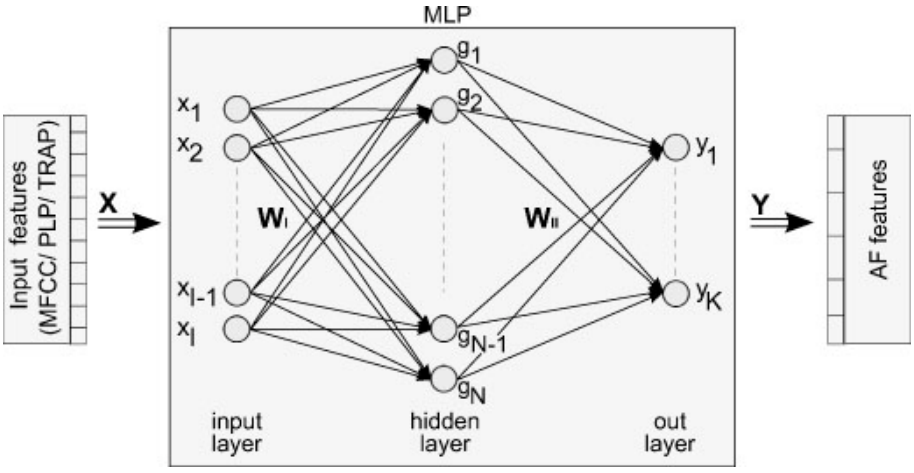

Fig. 1 *Structure of the three-layer MLP network.*

2.2 Neural network based estimation of articulatory features

There are several approaches which are used for an AF estimation, e.g. the classifiers based on the Gaussian Mixture Model, Support Vector Machine, Bayesian Networks, or the classifiers using multi-task learning which were used in [13, 30]. Nevertheless, the most widely used are the classifiers based on ANN [14, 38] that are very suitable especially in the situations where the classification should be done with the input acoustic speech feature vector of rather high dimension. This is exactly our case because we also work with the speech features with some context information.

For our AF classification, we used independent Multi-Layer Perceptrons (MLP) for particular AF classes. The MLP for each class has always the same 3-layer (I-N-K) structure, shown in Fig. 1. The input layer distributes the acoustic speech features, and therefore the number of its neurons depends on the size of an input feature vector (MFCC/PLP/TRAP). Each neuron output in the hidden layer is defined by a commonly used sigmoid activation function,

$$f_k(z_k) = \frac{1}{1 + e^{z_k}} \,. \tag{1}$$

An inner neuron potential z_k is computed in a standard way as a general weighted sum of the neuron inputs, i.e.

$$z_k = b_k + \sum_{j=1}^{I} w_{jk} x_j, \tag{2}$$

with the weights w_{jk} and bias b_k belonging to the k-th neuron. The MLP output represents a posteriori probability of given AF class with possible value in the range of $0 \div 1$. It is computed by a softmax activation function defined for the k-th output neuron and particular neuron potentials z_j as

$$f_k(z_k) = \frac{e^{z_k}}{\sum_{j=1}^{K} e^{z_j}}. \tag{3}$$

The size of the output layer is always given by the cardinality of the estimated AF class, as well as the size of the input layer is given by the size of the input speech features. The size of the hidden layer was set experimentally for the particular AF class and two different background environments.

2.3 Acoustic speech features for the AF estimation

All previously published works use a short context information on the input of the MLP networks in a different way, e.g. in the form of the first and the second derivatives of the features (delta and delta-delta features), by using the static features from more neighbouring short-time frames, or by using recurrent time-delay neural networks [2, 13, 14, 32]. A possible inclusion of the longer context on the input of the ANN-based AF classifier was discussed in [34].

In our work, *MFCC coefficients* were used as a general standard for the automatic speech recognition system in the following setup: 12 cepstral coefficients with the additional zeroth cepstral coefficient, 30 filters in the auditory-based spectral analysis, the frame length of 25 ms and frame step of 10 ms, preemphasis coefficient 0.97, and dynamic and accelerations features. The total length of this feature vector is 39. Similarly, *PLP cepstral coefficients* were used with the standard setup: 12 cepstral coefficients with the additional zeroth coefficient, 20 filters in the PLP-based auditory filter bank (for 16 kHz speech data), the frame length of 25 ms and frame step of 10 ms, and dynamic and accelerations features.

As a contribution on this study, we describe the usage of DCT-TRAP features for the AF estimation because these features were used successfully in very precise phoneme recognizers, e.g. [31], same as in other speech recognition applications. Our setup of the DCT-TRAP features was slightly modified. We used only 22 filters of the auditory spectral analysis, preemphasis coefficient 0.97, a short-time FFT frame length of 25 ms and the frame step of 10 ms, and a temporal pattern computed from 31 frames. This setup with a slightly shorter context (than it is used in [31]) and with the number of bands in the auditory spectral analysis smaller than within the standard MFCC computation was found to be a good compromise. It enables to decrease a computational complexity of MLP training due to the lower dimension of the speech feature vector and gives still sufficient accuracy of the AF estimation.

Concerning temporal patterns, they are represented by the Discrete Cosine Transform (DCT) to decrease both the input vector dimensionality and further decorrelation. In the end, 16 DCT coefficients were used which enabled to work finally with the dimension of $16 \times 22 = 352$ for the DCT-TRAP feature vector on the input of MLP.

3. Experiments

In the experimental part of this study, we first tested the basic accuracy of the AF estimation for Czech using three different previously chosen speech feature vectors and made a detailed analysis of the optimum setup of ANN, i.e. the optimum number of neurons in the hidden layer of the MLP network. In the second phase of the experiments we tested the robustness of this estimation for the speech collected under various conditions from the point of view of signal quality.

3.1 Experimental setup

All experiments were conducted with publicly available tools from *ICSI Quicknet Software Package* [12] (for ANN training and testing; currently newer TNet toolkit can be also used), from HTK Toolkit *HTK Toolkit* [37] (for speech features extraction and automatic phonetic segmentation), and also with our private tool *CtuCopy* [7] (for the feature extraction which supports a DCT-TRAP computation and *p-file* format required by the Quicknet toolkit).

The speech data for our experiments was taken from the Czech SPEECON database [4, 27] containing utterances from several environments collected by various microphones. We worked with two data sets, firstly with quite clean speech signals from the standard office environment (OFFICE subset) and secondly, with more noisy utterances from the car environment (a CAR subset). For these sets we have chosen the utterances with digits, and phonetically rich material (sentences and words) from all available records. Selected data was divided into three parts: non-overlapping training, cross-validation (CV), and test sets. The sizes of these sets from the OFFICE and CAR environments are summarized in more details in Tab. IV.

	OFFICE			CAR		
set	speakers	sentences	hours	speakers	sentences	hours
training	101	3450	4.99	48	4042	4.40
cross-val.	17	585	0.88	4	101	0.16
test	77	94	0.16	4	39	0.07

Tab. IV *OFFICE and CAR data subsets.*

The necessary phonetic segmentation of these sets was processed automatically by an HMM-based forced-alignment with available triphone acoustic models. For the test sets, phone boundaries were determined both automatically and manually. The manual phonetic segmentation enabled testing with the reference data with minimized error in the automatic phone boundary placement. The reference manual segmentation of the testing data was created according to the rules defined in [23] by engineers practised in phonetics and phonology. Some instances were also consulted with the experts from the Institute of Phonetics at Charles University, Prague.

Speech data was available from various input channels, so we were able to test also the robustness of the AF estimation for the same utterances collected with the microphones of different quality. In the end, we carried out the experiments with

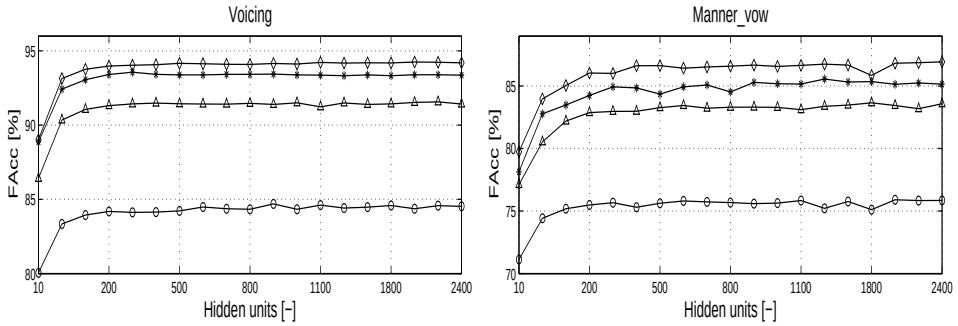

Fig. 2 *The number of hidden neurons in the MLP optimization for the DCT-TRAP feature; channel:* $\diamond$ *– CS0,* $*$ *– CS1,* Δ *– CS2,* $\circ$ *– CS3; environment: OFFICE.*

the signals collected by a high-quality head-set microphone (a close-talk channel CS0), by a medium quality basic hands-free microphone (a close-talk channel CS1), by high-quality microphones placed both in medium and far distances in the OFFICE environment (channels CS2 and CS3), and by other directional microphones in a medium distance in a CAR environment (channels CS2 and CS3). SNR levels strongly vary in these channels, from the average SNR about 26.82 dB (for channel CS0) representing a clean speech to the average SNR about 6.43 dB (for channel CS3) corresponding to the speech distorted by both convolutional and additive noise. SNR levels in all channels are presented in more details in Tab. V.

	CS0	CS1	CS2	CS3
OFFICE	26.82	19.51	12.71	6.43
CAR	14.00	6.95	12.06	8.99

Tab. V *Average values of SNR [dB].*

3.2 Results

The results were presented in percentage of the number of correctly recognized frames, i.e. the Frame Accuracy (*FAcc*) defined as

$$FAcc = \frac{n\_correct\_frame\_labels}{total\_frames} \cdot 100\,. \tag{4}$$

3.2.1 Optimal size of MLP

As mentioned above, we empirically looked for the optimal settings of MLP size for particular tasks, i.e. we analysed the *FAcc* of an AF estimation for $10 \div 2400$ neurons in the hidden layer.

The dependency of the frame accuracy on the number of hidden neurons is shown in Fig. 2 which contains the illustrative results for *Voicing* and *Manner_vow* estimation using DCT-TRAP on the input of MLP. The MLP sizes with the best

achieved accuracy of particular AF classifiers are summarized in Tab. VI, however, the optimal MLP sizes for individual AF classes could be lower, typically in the range of 300 ÷ 500 hidden neurons for *Voicing*, *Rounding*, and *Sonor*, and from 600÷800 for *Manner_vow*, *Place_vow*, *Manner_con*, and *Place_con* across all channels because the increment of *FAcc* for the higher numbers of hidden neurons lower is rather small but it also means higher computational costs. Finally, the same setup was used for both background conditions.

The computational time per one epoch of the MLP training depended on the size of the particular MLP and it was from 0.12 hours (for ANN with 10 neurons in the hidden layer) to 1.02 hours (for 2400 neurons in the hidden layer), for the experiments performed at PCs with CPU Intel(R) Core(TM) 2 Quad CPU Q9550 @ 2.83GHz with 4 cores and with CPU Intel(R) Core(TM) i7-2600 CPU @ 3.40GHz with8 cores. In comparison to [17], we did not use any special pre-processing for noisy data.

	Out units	*OFFICE*											
		Channel CS0				*Channel CS1*				*Channel CS2*			
		hids	*CV*	*Epoch*	*test*	*hids*	*CV*	*Epoch*	*test*	*hids*	*CV*	*Epoch*	*test*
Voicing	3	2000	94.9	4	94.3	300	94.2	4	93.6	2200	92.5	4	91.6
Place_con	9	1800	85.6	5	83.9	2200	83.8	5	82.0	1500	82.3	5	80.7
Place_vow	5	1500	88.5	5	88.3	1000	87.1	5	86.9	2200	85.5	5	85.2
Manner_con	9	1800	87.2	5	86.5	1800	85.6	5	84.2	2400	84.0	5	83.0
Manner_vow	5	2400	87.0	5	86.9	1300	85.8	5	85.6	1800	84.1	5	83.7
Rounding	4	1300	89.3	5	89.1	2400	88.0	5	87.8	1000	86.5	5	86.4
Sonor	4	1300	89.0	5	88.9	2200	88.0	5	87.5	1500	86.1	5	85.8
		CAR											
Voicing	3	1000	94.6	4	87.0	1300	92.9	3	85.3	2200	91.9	4	83.9
Place_con	9	1500	85.6	8	77.8	2400	83.0	7	75.7	2000	82.0	7	74.6
Place_vow	5	2200	88.8	4	80.1	1500	87.2	5	79.2	500	86.2	5	77.7
Manner_con	9	2400	86.2	7	79.1	2000	84.0	6	77.0	1500	82.9	7	75.7
Manner_vow	5	1800	87.5	4	79.0	2200	86.3	5	78.4	2400	84.8	5	77.0
Rounding	4	2000	89.4	5	81.3	2000	87.6	5	80.4	2000	86.5	5	78.8
Sonor	4	1000	87.9	6	80.8	2000	86.2	4	78.7	2200	85.5	6	78.0

Tab. VI *Best setup size of MLP for the best results with DCT-TRAP features.*

3.2.2 Robustness of an MLP-based AF estimation

The previously published results obtained by other authors for English presented in [5] can be summarized in the following numbers:

- *voicing*: average accuracy 90.3 %, the best 93.0 %,
- *place*: average 75.4 %, the best 85.9 %,
- *manner*: average 85.3 %, the best 88.5 %,
- *rounding*: average 86.2 %, the best 92.0 %,
- *front-back*: average 83.7 %, the best 87.4 %.

These values, however, were not always obtained under the absolutely same conditions. Some authors also measured the accuracy at a phone level, others at a frame level (as in our case), so the exact comparison is difficult. Also we distinguished between the vowels and consonants for the manner and place of articulation.

Our results obtained for the OFFICE environment are summarized in Fig. 3. These results for the MFCC and PLP features proved a reliable standard estimation of AFs for Czech, which is comparable to the results of other authors. Our best results for all AF classes were obtained for DCT-TRAP features. For a high-quality CS0 channel it was 94.3% for the classiffication of the voicing class, 83.9% for the place of a consonant, 88.3% for the place of a vowel, 86.5% for the manner of a consonant, 86.9% for the manner of a vowel, 89.1% for rounding, and 88.9% for sonoring. As concerns individual AF classes, the best results were obtained for the voicing detection, whereas the most difficult seemed to be the estimation of the place of articulation for the consonants. Using DCT-TRAP features, only slightly worse results were obtained for other more noisy channels (CS1, CS2 and CS3).

The results achieved in the evaluations with manually-labelled reference data are presented in Tab. VII using the average *FAcc* (calculated across AFs). Better results were always achieved in the evaluations with automatically labelled reference data. The AF evaluation with manually-labelled reference data represented mismatched conditions because the data with automatic phonetic segmentation were used for the training. However, these results have the similar trend as those obtained by the reference data labelled automatically.

The results for a more noisy CAR environment, see Fig. 4 and Tab. VII, prove the robustness of an MLP-based AF estimation, especially, when the DCT-TRAP features were used as an output of the acoustic analysis. We can see only small decrease in *FAcc* in comparison to the results obtained for clean speech data from the OFFICE environment. However, when comparing the results from these two environments, we must note that channels CS2 and CS3 contain the speech of slightly different quality.

	MFCC				PLP				DCT-TRAP			
	CS0	CS1	CS2	CS3	CS0	CS1	CS2	CS3	CS0	CS1	CS2	CS3
OFFICE	81.4	79.7	78.4	73.6	81.6	79.9	78.6	73.5	82.3	81.3	80.3	74.5
CAR	85.2	83.6	81.3	81.6	85.3	83.4	81.4	81.6	85.0	85.3	83.8	83.5

Tab. VII *Average FAcc of an AF estimation for the manually labelled data.*

The robustness of the MLP-based AF estimation was also observed when the training and testing conditions were not the same, since it is a common situation in real deployed systems having a significant influence on the speech recognition accuracy [29]. These analyses are presented using the average *FAcc* trend (average was computed across all AFs) and the results for the channel mismatch and environment mismatch are summarised in Fig. 5. The impact of switching from the close-talk microphone to the far-talk one is presented in Figs. 5A and 5B. The robustness of a DCT-TRAP AF estimation is demonstrated by very small decrease in an average *FAcc*, when the training was performed on the CS0 channel only, especially in case of the CAR environment. The highest decrease was observed

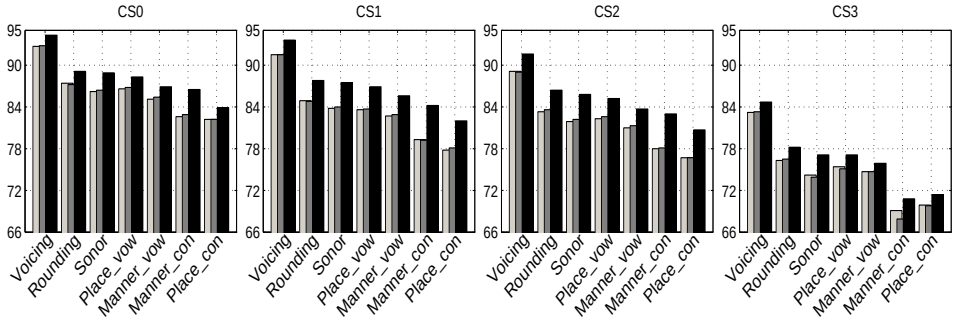

Fig. 3 *FAcc of an AF estimation for an automatically labelled OFFICE test set; features: MFCC – light gray, PLP – dark gray, DCT-TRAP – black.*

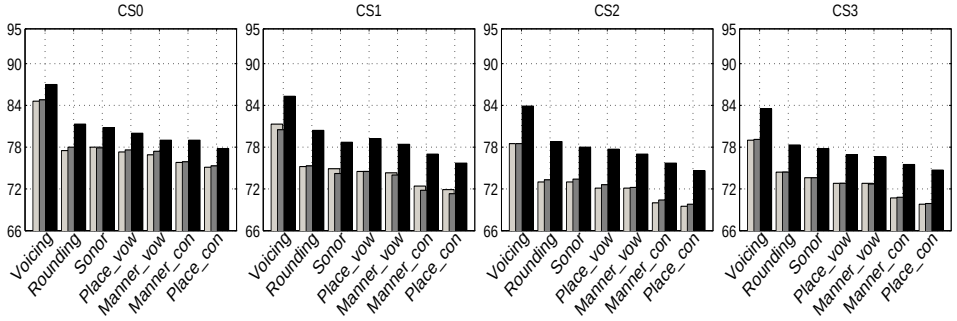

Fig. 4 *FAcc of an AF estimation for an automatically labelled CAR test set; features: MFCC – light gray, PLP – dark gray, DCT-TRAP – black.*

for the CS3 channel in the OFFICE environment, however, it was the result when significantly degraded speech with SNR of about 6 dB only was analysed. An environmental mismatch represents higher influence on the reached AF classification accuracy, see in Fig. 5C. The decrease in average *FAcc* of an AF estimation in the CAR environment was about $6 \div 9\%$ for individual channels, when the training was performed on the OFFICE data.

	OFFICE					*CAR*				
	CS0	*CS1*	*CS2*	*CS3*	*avg*	*CS0*	*CS1*	*CS2*	*CS3*	*avg*
Voicing	94.3	93.6	91.6	84.7	*91.1*	87.0	85.3	83.9	83.5	*84.9*
Place_con	83.9	82.0	80.7	71.4	*79.5*	77.8	75.7	74.6	74.7	*75.7*
Place_vow	88.3	86.9	85.2	77.1	*84.4*	80.0	79.2	77.7	76.9	*78.5*
Manner_con	86.5	84.2	83.0	70.8	*81.1*	79.0	77.0	75.7	75.5	*76.8*
Manner_vow	86.9	85.6	83.7	75.9	*83.0*	79.0	78.4	77.0	76.6	*77.8*
Rounding	89.1	87.8	86.4	78.2	*85.4*	81.3	80.4	78.8	78.3	*79.7*
Sonor	88.9	87.5	85.8	77.1	*84.8*	80.8	78.7	78.0	77.8	*78.8*

Tab. VIII *FAcc of DCT-TRAP AF estimation for speech degraded by car noise.*

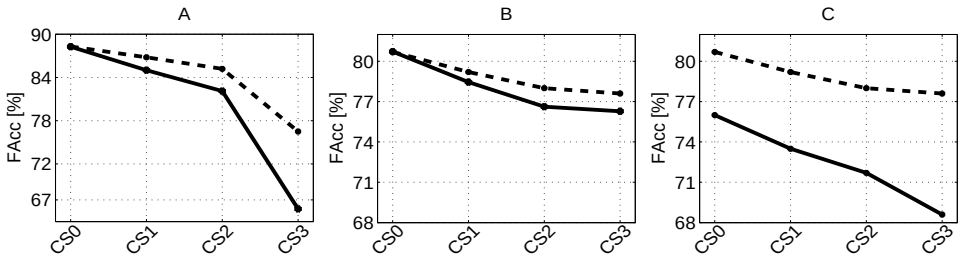

Fig. 5 *Average FAcc (across all AFs) for mismatched condition: A – channel mismatch in OFFICE environment (dashed line – matched training, solid line – training on CS0); B – channel mismatch in CAR environment (dashed – matched training, solid – training on CS0); C – environmental mismatch in CAR environment (dashed – matched training in CAR; solid – training on OFFICE data).*

Finally, the decrease in *FAcc* for the individual AFs influenced by the car noise is presented in Tab. VIII. As regards the individual AFs, the decrease was about $4 \div 6\%$ only when DCT-TRAP acoustic features were used on the input of MLPs.

4. Conclusions

In this paper, we analysed the MLP-based estimation of AFs and we proposed its robust approach based on DCT-TRAP acoustic features. One of the contributions of this paper lies in creating the basic design of the computation of the articulatory features for the Czech speech. The basic classes of AF features were defined in the same way as defined for English, with respect to the specific peculiarities for Czech.

Secondly, the fundamental study of the MLP-based AF estimation accuracy for the Czech speech was carried out. The optimal setups of MLPs for all AFs and all analysed acoustic inputs were found, and the achieved results for the best MLP setup under clean conditions were compared with the results for English published by other authors. We have achieved similar values as average results as it was published for English. After increasing the size of the hidden layer and after using the DCT-TRAP features, almost the same accuracies were obtained also for the AF estimation in the car noise conditions.

Finally, this approach using DCT-TRAPs seems to be a very good and robust way of the AF estimation, so the combination of the standard features and AFs obtained by this approach is supposed to have a positive influence on the accuracy of the recognition or automatic phonetic segmentation of degraded speech [11].

Within further work we suppose to perform the optimization of AF estimation. For this purpose the modern approaches based on DNN or articulatory bottleneck features are supposed to be used. AFs should be then used for AF-based phone recogniton, the phonetic segmentation, and AF-TANDEM-based ASR as particular steps to the target spontanoeus and informal speech recognition with the special focus on Czech.

Acknowlegment

Research described in the paper was supported by an internal CTU grant SGS14/191/OHK3/3T/13 "Advanced Algorithms of Digital Signal Processing and their Applications".

References

[1] CHANG S., GREENBERG S., WESTER M. An elitist approach to articulatory-acoustic feature classification. In: *Proceedings of the 7th European Conference on Speech Communication and Technology (Eurospeech 2001 – Scandinavia)*, Aalborg, Denmark. Aalborg: ISCA, 2001, pp. 1725–1728.

[2] CHANG S., WESTER M., GREENBERG S. An elitist approach to automatic articulatory-acoustic feature classification for phonetic characterization of spoken language. *Speech Communication*. 2005, 47(3), pp. 290–311, doi: 10.1016/j.specom.2005.01.006.

[3] DENG L., HINTON G., KINGSBURY B. New types of deep neural network learning for speech recognition and related applications: an overview. In: *Proceedings of the 2013 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2013)*, Vancouver, Canada. Vancouver: IEEE, 2013, pp. 8599–8603, doi: 10.1109/ICASSP.2013.6639344.

[4] ELRA. Czech SPEECON adult database [software]. 2009-03-23 [accessed 2014-09-18]. Available from: http://catalog.elra.info/product_info.php?products_id=1095

[5] FRANKEL J., WESTER M., KING S. Articulatory feature recognition using dynamic Bayesian networks. *Computer Speech & Language*. 2007, 21(4), pp. 620–640, doi: 10.1016/j.csl.2007.03.002.

[6] FOUSEK P. *Extraction of Features for Automatic Recognition of Speech Based on Spectral Dynamics* [online]. Prague, 2007. PhD thesis, Czech Technical University in Prague [viewed 2014-09-18]. Available from: http://noel.feld.cvut.cz/speechlab/publications/051_disertace07.pdf

[7] FOUSEK P., et al. CtuCopy – Universal feature extractor and speech enhancer [software]. 2013-06-13 [accessed 2014-09-18]. Available from: http://noel.feld.cvut.cz/speechlab

[8] GRÉZL F. *TRAP-based probabilistic features for automatic speech recognition* [online]. Brno, 2007. PhD thesis, Brno University of Technology [viewed 2014-09-18]. Available from: http://www.fit.vutbr.cz/research/view_pub.php?id=8518

[9] GRÉZL F., FOUSEK P. Optimizing bottle-neck features for LVCSR. In: *Proceedings of the 2008 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2008)*, Las Vegas, Nevada, USA. Las Vegas: IEEE, 2008, pp. 4729–4732, doi: 10.1109/ICASSP.2008.4518713.

[10] HERMANSKY H. Perceptual linear predictive PLP analysis of speech. *Journal of the Acoustic Society of America*. 1990, 87(4), pp. 1738–1752, doi: 10.1121/1.399423.

[11] HOSOM J.P. Automatic phoneme alignment based on acoustic-phonetic modeling. In: *Proceedings of the 7th International Conference On Spoken Language Processing (ICSLP 2002)*, Denver, Colorado, USA. Denver: ISCA, 2002, pp. 357–360.

[12] JOHNSON D., et al. ICSI Quicknet Software Package [software]. 2004-02-06 [accessed 2014-09-18]. Available from: http://www.icsi.berkeley.edu/Speech/qn.html

[13] KING S., et al. Speech production knowledge in automatic speech recognition. *Journal of the Acoustic Society of America*. 2007, 121(2), pp. 723–742, doi: 10.1121/1.2404622.

[14] KING S., TAYLOR P. Detection of phonological features in continuous speech using neural networks. *Computer Speech & Language*. 2000, 14(4), pp. 333–353, doi: 10.1006/csla.2000.0148.

[15] KINGSBURY B.E.D., MORGAN N., GREENBERG S. Robust speech recognition using the modulation spectrogram. *Speech Communication*. 1998, 25(1-3), pp. 117–132, doi: 10.1016/S0167-6393(98)00032-6.

[16] KIRCHHOFF K. Combining articulatory and acoustic information for speech recognition in noisy and reverberant environments. In: *Proceedings of the 5th International Conference On Spoken Language Processing (ICSLP 1998)*, Sydney, Australia. Sydney: ISCA, 1998, pp. 891–894.

[17] KIRCHHOFF K. *Robust speech recognition using articulatory information* [online]. Bielefeld, 1999. PhD thesis, Technische Fakultät der Universität Bielefeld [viewed 2014-09-18]. Available from: http://pub.uni-bielefeld.de/download/2302713/2302716

[18] KIRCHHOFF K., FINK G.A., SAGERER G. Combining acoustic and articulatory feature information for robust speech recognition. *Speech Communication*. 2002, 37(3-4), pp. 303–319, doi: 10.1016/S0167-6393(01)00020-6.

[19] LAL P., KING S. Cross-lingual automatic speech recognition using tandem features. *IEEE Trans. on Audio, Speech, and Language Processing*. 2013, 21(12), pp. 2506–2515, doi: 10.1109/TASL.2013.2277932.

[20] LIVESCU K., et al. *Articulatory feature-based methods for acoustic and audio-visual speech recognition: 2006 JHU summer workshop final report* [online]. Baltimore: Johns Hopkins University, 2006 [viewed 2014-09-18]. Research report. Available from: http://www.isle.illinois.edu/sst/pubs/2006/WS06AFSR_final_report.pdf

[21] LIVESCU K., et al. Manual transcription of conversational speech at the articulatory feature level. In: *Proceedings of the 2007 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2007)*, Honolulu, USA. Honolulu: IEEE 2007, pp. 953–956, doi: 10.1109/ICASSP.2007.367229.

[22] LIVESCU K., FOSLER-LUSSIER E., METZE F. Subword modeling for automatic speech recognition: Past, present, and emerging approaches. *IEEE Signal Processing Magazine*. 2012, 29(6), pp. 44–57, doi: 10.1109/MSP.2012.2210952.

[23] MACHAČ P., SKARNITZL R. *Fonetická segmentace hlásek*. Prague: Epocha, 2009.

[24] METZE F. *Articulatory Features for Conversational Speech Recognition* [online]. Karlsruhe, 2005. PhD thesis, Universität Fridericiana zu Karlsruhe [viewed 2014-09-18]. Available from: http://digbib.ubka.uni-karlsruhe.de/volltexte/documents/3266

[25] MIKOLOV T., ZWEIG G. Context dependent recurrent neural network language model. In: *Proceedings of the 2012 IEEE Workshop on Spoken Language Technology (SLT 2012)*, Miami, Florida, USA. Miami: IEEE, 2012, pp. 234–239, doi: 10.1109/SLT.2012.6424228.

[26] PALKOVÁ Z. *Fonetika a fonologie češtiny*. Prague: Karolinum, 1997.

[27] POLLÁK P., ČERNOCKÝ J. *Czech SPEECON adult database* [online]. Prague, 2004. Technical report, Czech Technical University in Prague [viewed 2014-09-18]. Available from: http://noel.feld.cvut.cz/speechlab/publications/094_speecon_design.pdf

[28] PRABHAVALKAR R., et al. Discriminative articulatory models for spoken term detection in low-resource conversational settings. In: *Proceedings of the 2013 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2013)*, Vancouver, Canada. Vancouver: IEEE, 2013, pp. 8287–8291, doi: 10.1109/ICASSP.2013.6639281.

[29] RAJNOHA J., POLLAK P. ASR systems in noisy environment: Analysis and solutions for increasing noise robustness. *Radioengineering*. 2011, 20(1), pp. 74–84, Available from: http://www.radioeng.cz/fulltexts/2011/11_01_074_084.pdf

[30] RASIPURAM R., MAGIMAI-DOSS M. *Multitask learning to improve articulatory feature estimation and phoneme recognition* [online]. Martigny: IDIAP, 2011 [viewed 2014-09-18]. Research report Idiap-RR-21-2011. Available from: http://publications.idiap.ch/index.php/publications/show/2103

[31] SCHWARZ P. *Phoneme Recognition based on Long Temporal Context* [online]. Brno, 2009. PhD thesis, Brno University of Technology [viewed 2014-09-18]. Available from: http://www.fit.vutbr.cz/∼schwarzp/publi/thesis.pdf

[32] SINISCALCHIA S.M., LEEB C. A study on integrating acoustic-phonetic information into lattice rescoring for automatic speech recognition. *Speech Communication*. 2009, 51(11), pp. 1139–1153, doi: 10.1016/j.specom.2009.05.004.

[33] TUČKOVÁ J., KOMÁREK V. Effectiveness of speech analysis by self-organizing maps in children with developmental language disorders. *Neuroendocrinology Letters*. 2008, 29(6), pp. 939–948.

[34] VALENTE F., DOSS M.M., WANG W. Analysis and comparison of recent MLP features for LVCSR systems. In: *Proceedings of the 12th Annual Conference of the International Speech Communication Association (Interspeech 2011)*, Florence, Italy. Florence: ISCA, 2011, pp. 1245–1248.

[35] VOLÍN J. Fonetika a fonologie. In: CVRČEK V., et al., eds. *Mluvnice současné češtiny*. Prague: Karolinum, 2013, pp. 105–145.

[36] WELLS J.C., et al. Czech SAMPA Home Page. In: *SAMPA – computer readable phonetic alphabet* [online]. University College London, 2003 [viewed 2014-09-18]. Available from: http://www.phon.ucl.ac.uk/home/sampa/czech-uni.htm

[37] YOUNG S., et al. *The HTK Book, Version 3.4.1* [online]. Cambridge University Press, 2009 [viewed 2014-09-18]. Available from: http://htk.eng.cam.ac.uk

[38] YU D., et al. Boosting attribute and phone estimation accuracies with deep neural networks for detection-based speech recognition. In: *Proceedings of the 2012 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2012)*, Kyoto, Japan. Kyoto: IEEE, 2012, doi: 10.1109/ICASSP.2012.6288837.