Analysis and Optimization of Telephone Speech Command Recognition System Performance in Noisy Environment

Jan NOVOTNÝ, Pavel SOVKA, Jan UHLÍŘ

Dept. of Circuit Theory, Czech Technical University, Technická 2, 166 27 Praha, Czech Republic

novotnj2@feld.cvut.cz, sovka@feld.cvut.cz, uhlir@feld.cvut.cz

Abstract. *This paper deals with the analysis and optimization of a speech command recognition system (SCRS) trained on Czech telephone database Speechdat(E) for use in a selected noisy environment. The SCRS is based on hidden Markov models of context dependent phones (triphones) and mel-frequency cepstral coefficients analysis of speech (MFCC). The main aim is to analyze and to search for the optimal settings of SCRS with respect to additive noise robustness without use of additional techniques for additive noise reduction. The analysis is pointed to the appropriate setting of MFCC computation, the silence model adjustment and grammar selection possibilities. It is shown, that the correct performance of SCRS strictly depends on an appropriate adjustment of the silence model. The ability of the silence model adaptation is confirmed. When SNR is higher than 15 dB the suitable performance of SCRS can be guarantied without any modification of the triphones speech models by: 1. the optimal setting of MFCC computation, 2. the proper silence model adaptation. The assumption of a speech command recognition system use in an environment where SNR is higher than 15 dB is fulfilled in many applications.*

Keywords

Robust speech recognition, Mel-cepstral analysis, silence model adaptation, parallel model combination.

1. Introduction

The great effort has been devoted to the development of noise robust speech recognition systems [1]-[5]. The main aim of this contribution is the analysis, limit finding and optimization of the telephone speech command recognition system (SCRS) performance in a selected noisy environment without use of additional techniques for additive noise reduction [2], [5]. The speech recognition system analysis is divided into three parts. The first deals with the optimal setting of MFCC computation with respect to the additive noise presence. The second addresses the influence of the silence model parameters setting. Two possible grammar constructions for SCRS are compared in the last part. The deeper understanding of SCRS performance degradation by the influence of an additive noise is the motivation for all this work.

2. Speech Recognition System Construction

The context dependent hidden Markov models (HMM) of phonemes trained on two thirds of Czech database for the fixed telephone network Speechdat(E) [6] were used for the speech recognition system building. The whole database consists of approximately 100 hours of speech records from 1000 speakers. This type of SCRS construction enables an undemanding configuration of commands dictionary in comparison to SCRS based on HMM models of full words. Mel-frequency cepstral coefficients analysis, which is frequently being applied as a base for noise robust front-ends was selected [13], [14].

2.1 The Testing Database and Noises Selection

One tenth of database Speechdat(E) [6] was used for testing. This part was selected with the intention not to overlap with the training part of the database. The testing database contains speech records with ten Czech numerals zero to nine in random order. The pauses between testing words have random duration, which well simulates real applications.

The analysis of the noise naturally present in the telephone communication was performed [12]. It was found that the naturally occurring noise can be separated by its characteristics into three groups. The noise from the first group (*Noise*$_1$) has stationary character and is mainly caused by the transmission channel operation. The most common value of *SNR* is between 35 to 50 dB for this type of noise. This noise is present both in pauses and during the speech activity. The second category noise (*Noise*$_2$) is caused by speaker breathing; it has medium-term duration (0.5–1 s) and a characteristic spectrum. The most common value of *SNR* is between 15 to 30 dB for this type of noise. The important fact is that this type of noise can be found just in pauses and it does not affect the speech. The last group (*Noise*$_3$) represents the noise, which is caused by manipulation with an earpiece by the speaker. This type of noise has short-term duration with relatively high level and is also often registered during pauses. All types of noise are commonly present during HMM training so SCRS is well adapted on them and is able to operate with recognition results higher than 95 %. The situation when the noise, which was not included during training stage, is simulated below. In the real application it would be the noise of an air-conditioner, a computer fan noise, a printer noise etc.

A synthetically generated stationary white noise and three types of stationary narrowband noises were used for the robustness of SCRS against the additional additive noises testing. The three types of stationary narrowband noises were generated by white noise filtering in purpose to affect the first (noise F_1, frequency band between 0.3 and 0.9 kHz), the second (noise F_2, frequency band between 1 and 2.5 kHz) or the third (noise F_3, frequency band between 2.5 and 3.4 kHz) formants. These noises are added to the testing speech records in order to achieve the specified *SNR* according to

$$x[k]=s[k]+\sqrt{\left(\frac{\hat{P}_s}{\hat{P}_n}\right)10^{-\frac{SNR}{10}}}\,n[k], \qquad (1)$$

where

$x[k]$ is the output signal,

$s[k]$ is the original speech,

$n[k]$ represents testing noise,

$\hat{P}_s$ is the power of speech (obtained with use of forced alignment),

$\hat{P}_n$ is the power of the testing noise.

2.2 HMM Parameters and Training Method Description

The observation vector is composed of three streams. The first stream is represented by 12 static mel-cepstral coefficients and one energy coefficient. The second and the third stream are composed of delta and acceleration coefficients respectively (parameterization MFCC_E_D_A [15]). During some experiments the energy coefficient is replaced by a 0'th cepstral (C_0) coefficient or it is not used at all. If it is done in this way, it is emphasized in the text. The context-dependent phonemes are modeled as a three-state HMM with tied states. Tree-based clustering was performed. The HMM training method is almost identical with [15], only the output distributions of all states are divided into three streams (as mentioned before) and each stream into three-component Gaussian mixture and three reestimations are added at the end of HMM training. Two silence models with the identical structure like [15] are used.

3. The SCRS Performance under the Influence of Additive Noise

Extensive experiments were carried out with the objective to analyze the SCRS performance under the influence of additive noise. The first intention is to obtain good recognition results at relatively high signal to noise ratio. The second intention is to obtain comparable results both for parameter *acc* (Percent Accuracy [%]) and for parameter *corr* (Percent Correct [%]) [15]. The parameter *corr* definition does not account for insertion errors [15], in this case for extra inserted words. The extra inserted words are most often inserted during pauses. This is why the difference between *acc* and *corr* parameters can be found useful for SCRS performance during pauses evaluation. If both parameters show similar results then it can be supposed that the extra inserted command error is suppressed.

3.1 Mel-cepstral Analysis Setting Influence

The recognition system robustness against the artificially added noise with respect of MFCC setting is described. The appropriate settings of *M* = WINDOWSIZE (the time duration of an input speech frame), *tr* = TARGETRATE (time shift between two subsequent speech frames) and θ = DELTAWINDOW = ACCWINDOW (the number of subsequent frames used for delta and acceleration coefficients computation) [15] parameters are investigated for parameterization MFCC_E_D_A [15]. When particular optimum of the previous parameters is found, the importance of the energy coefficient and its possible replacement by the 0'th cepstral coefficient is evaluated.

A new set of HMM for every tested setting was trained. The recognition system performance with the obtained models was tested with use of all defined artificial noises within the *SNR* interval 0-40 dB. The analysis of experiments proved that the *M*, *tr* and θ parameters setting significantly influences the recognition system robustness against the additive noise. Tab. 1 shows the dependence of recognition system robustness against white noise on *tr* parameter. The best results were obtained for *tr* = 16 ms (the third column of Tab. 1.). When smaller value of *tr* parameter is set (10 ms) and the analyzed window size is held on *M* = 32 ms then the recognition system robustness can be increased again by the modification of delta and acceleration coefficients computation. This means that these coefficients are computed from longer signal segment (see Tab. 2). This fact confirms the idea that the *tr* setting and the way of delta and acceleration coefficients computation are related. It means that for shorter *tr* time it is valuable to compute the delta and acceleration coefficients from wider time surroundings of current segment. The appropriate choice of *tr*, *M* and θ parameters improves the SCRS performance during pauses in speech which can be seen from the difference between *corr* and *acc* results. This fact relates with better determination of pauses by the silence model.

The importance of the energy coefficient is evaluated in Tab. 3. There are three dependences of recognition results on the *SNR*, which were measured with the energy coefficient use (without normalization of the energy coefficient), the C_0 coefficient use and when no one of them was used, respectively. The results obtained with energy coefficient are slightly better than with C_0 coefficient. The recognition results are significantly worse if no one of them was used.

Just the white noise dependences were selected (as typical) to document the previous experiments. The results for all types of defined testing noises and parameters $tr =$ $= 16$ ms, $M = 32$ ms a $\theta = 2$ are shown in Tab. 4 and will be referred to in the next sections. The best recognition results were obtained for white noise and the worst for F_3 noise generally in all these simulations. Thus F_3 noise can be identified as the most harmful from this point of view.

SNR	white noise, tr =			
	8 ms	10 ms	16 ms	20 ms
[dB]	Corr/Acc	Corr/Acc	Corr/Acc	Corr/Acc
30	95.5/90.3	97.1/94.8	97.5/97.1	95.4/95.2
24	92.8/80.7	95.4/90.9	96.5/96.1	94.2/94.2
18	86.3/56.1	90.9/75.2	93.4/92.7	92.5/92.5
12	68.3/33.5	69.8/48.7	83.4/81.4	82.4/82.4
6	37.7/23.8	33.3/24.9	50.3/49.9	52.4/52.0
0	12.0/11.2	11.0/10.8	14.1/13.9	20.5/20.3

Tab. 1. SCRS robustness dependence on *tr* ($M = 2$ *tr*, $\theta = 2$).

SNR	white noise, θ =			
	2	3	4	5
[dB]	Corr/Acc	Corr/Acc	Corr/Acc	Corr/Acc
30	97.7/94.2	96.5/95.0	96.7/96.1	95.7/95.2
24	95.9/89.9	95.0/92.8	95.4/94.2	94.0/93.8
18	89.6/68.9	89.6/80.7	91.1/89.8	91.9/91.5
12	70.6/43.1	73.5/52.6	77.0/68.1	78.9/77.8
6	36.6/26.1	38.9/26.7	45.5/35.8	54.2/52.4
0	11.4/11.0	13.5/13.0	14.7/13.9	23.2/22.6

Tab. 2. SCRS robustness dependence on θ (tr = 10 ms, M = 32 milliseconds).

3.2 Silence Model Setting Influence

The experiments carried out and their analysis proved that if the additive noise is present the SCRS performance is often getting worse during pauses [10], [11]. This can be found by the analysis of recognized commands and their time alignment; indirectly it can be detected by the *acc* and *corr* parameters difference. It was also found that the type of the most often incorrectly inserted command is dependent on the additive noise type used for testing. Because the SCRS performance during pauses is closely related to the silence model, this model setting influence on SCRS performance in additive noise was tested.

SNR	white noise, parameters:		
	E	C0	-
[dB]	Corr/Acc	Corr/Acc	Corr/Acc
30	97.5/97.1	96.9/96.7	96.7/96.1
24	96.5/96.1	95.7/95.5	94.0/93.2
18	93.4/92.7	91.5/90.9	85.5/83.8
12	83.4/81.4	78.5/78.3	59.0/57.8
6	50.3/49.9	43.9/43.3	25.1/24.6
0	14.1/13.9	9.9/9.9	10.6/10.6

Tab. 3. SCRS robustness dependence for white noise when the E or C0 or no additional parameterization coefficient is used (tr = 16 ms, M = 32 ms, θ = 2).

SNR	white noise	F_1 noise	F_2 noise	F_3 noise
[dB]	Corr/Acc	Corr/Acc	Corr/Acc	Corr/Acc
40	98.1/96.7	97.9/96.3	97.9/96.7	98.1/96.1
30	98.1/97.5	98.1/95.0	97.5/95.7	95.9/89.2
25	96.9/95.9	97.1/89.6	95.7/92.5	91.3/80.1
20	95.5/94.8	91.3/75.4	88.2/82.2	83.0/67.3
15	93.0/91.7	80.5/62.1	70.4/64.4	73.5/56.3
10	80.1/77.8	60.1/44.7	47.4/44.7	58.4/44.9
5	46.6/45.3	33.9/26.9	25.5/24.0	48.5/36.6
0	13.0/13.0	18.2/16.1	14.1/13.5	29.6/26.5

Tab. 4. Reference results: tr = 16 ms, M = 32 ms and θ = 2.

The silence model differs from the other speech models. For example, if SCRS operated in an environment without noise then the silence model would correspond to a signal with very low energy and random spectrum. If SCRS operates in slightly noisy environment (almost in any real application) then the silence model represents this environment. If the environment varies then the silence model should also vary, because otherwise it is not corresponding to the environment and SCRS is unable of correct pauses (silence) identification.

A three state silence model with forward-backward skip between the first and the last state [15] was used during all previous simulations. This model represents both short and long duration pauses. This silence model also represents the noise naturally present in telephone communication. The test database selection corresponds to the reality, because the speakers had not been given the instructions to make equal time pauses between commands (numbers) and naturally present noise had not been removed in any way.

The silence model parameters settings and its influence to the SCRS performance in additive noise is investigated in the next sections. The SCRS performance is tested with the retrained silence model at first. This silence model was retrained in the way to be corresponding to one type of testing artificial noise. The next attempt is to study of

SCRS performance when the silence model setting is being dynamically changed to be matched to the actually tested noise.

SNR	white noise	F_1 noise	F_2 noise	F_3 noise
[dB]	Corr/Acc	Corr/Acc	Corr/Acc	Corr/Acc
40	92.7/83.6	92.7/83.6	92.7/83.6	92.7/83.6
30	90.5/78.1	92.1/83.2	88.0/80.3	96.9/94.8
25	87.8/72.3	90.7/79.7	81.4/76.0	95.5/95.2
20	80.3/62.3	84.3/71.0	66.5/63.2	92.3/91.9
15	67.9/55.7	71.8/60.1	42.5/40.0	87.4/87.4
10	48.0/41.2	46.6/40.2	25.1/24.0	80.5/80.5
5	14.9/14.3	24.2/21.1	15.1/14.7	56.9/56.3
0	10.2/10.2	14.5/13.9	12.2/12.0	27.5/27.3

Tab. 5. Recognition rates for the retrained silence model to be matched F_3 and *SNR* = 20 dB conditions (Parameterization MFCC_E_D_A, *tr* = 16 ms, *M* = 32 ms, θ = 2).

3.2.1 The Influence of Silence Model Obtained by Retraining

The set of HMM was trained on two thirds of SpeechDat(E) database (the results of these models set are presented in Tab. 4). The silence model obtained by this training was replaced by the silence model which is more representative for hypothetical operation conditions. These conditions are specified by F_3 noise occurrence with *SNR* of 20 dB. The new silence model was created by standard training procedure on small part of SpeechDat(E) database (containing approximately one hundred of records) to which the F_3 noise with 20 dB *SNR* was artificially added. The analyses of SCRS performance with retrained model of noise were carried out in standard way for four types of testing noise and *SNR* range between 0 dB and 40 dB. The results are presented in Tab. 5. From the comparison of recognition results for the original silence model (Tab. 4) and for the retrained silence model (Tab. 5) the following can be seen: The significant improvement of recognition results was achieved in conditions which are in accordance to the silence model (the dependence for F_3 noise). On the contrary if the testing noise does not correspond to the silence model, the SCRS performance becomes significantly worse. On condition that the noise parameters and the silence model parameters are similar the *corr* and *acc* results come close. This fact gives the evidence of good SCRS performance during pauses.

If we liked to carry out such analysis of SCRS performance with matched silence model, the silence model would need to be retrained again for all tested types of noise and *SNR*. This method is possible, but it is computationally expensive and time consuming, furthermore the silence model parameters obtained by model training can be hardly analyzed. Therefore two different ways of the silence model parameters estimation were used.

3.2.2 The Influence of Adapted Silence Model

A simple estimation method of silence model parameters was proposed in purpose of SCRS performance analysis with adapted silence model. Both testing noise and naturally present noise are considered within silence model computation. The silence model parameters are directly estimated from corresponding testing noise realizations (white noise, F_1, F_2 and F_3 noise) with addition of selected naturally present noise realizations (*Noise*$_1$), (*Noise*$_2$) and (*Noise*$_3$). The noise obtained in this way is parameterized into vectors o_t. The mean vector $\hat{\mu}_j$ and the diagonal covariance matrix $\hat{\Sigma}_j$ are estimated by following equations

$$\hat{\mu}_j = \frac{1}{T}\sum_{t=1}^{T} o_t, \tag{2}$$

$$\hat{\Sigma}_j = \frac{1}{T}\sum_{t=1}^{T} (o_t - \mu_j)(o_t - \mu_j)'. \tag{3}$$

A three state silence model is built from the estimated parameters. All states are identical and the original transition matrix is used. The same three-component Gaussian mixtures are used in all three states. The noise parameters are estimated for every mixture separately with intention to include the different type of naturally present noise to every mixture. In other words, the sum of testing noise and naturally present noise of *Noise*$_1$ type was used for mixture 1 parameters estimation. The *Noise*$_2$ noise was added in mixture 2 parameters estimation and the *Noise*$_3$ noise was added in mixture 3 parameters estimation. The final results of the SCRS performance with adapted silence model are presented in Tab. 6.

3.2.3 The Influence of Log-Add PMC Adapted Silence Model

Parallel model combination (PMC) method based on Log-Add approximation [8]-[9] was tested with purpose of silence model adaptation to the testing noises. The Log-Add approximation can be seen as a simplification of an original Log-Normal [7] approximation and it makes the assumption that the HMM models to be compensated have zero variance. Nevertheless a very good performance was reported [8]. The main disadvantage of PMC based methods is a high computational load when all the HMM models have to be compensated. It is not true in this case when just the silence model is compensated. The whole process is described by following equation

$$\hat{\mu}_j^l = \log\left(\exp(\mu_j^l) + \exp(\tilde{\mu}_j^l)\right), \tag{4}$$

where $\hat{\mu}_j^l$, μ_j^l and $\tilde{\mu}_j^l$ are the new silence model, the original (trained) silence model and the tested noise means in the Log-Spectral domain. The features in cepstral domain are transformed into log-energy domain and back via the discrete cosine transform (DCT). To be able to perform

DCT correctly the energy parameter in feature vector is replaced by C_0 coefficient. The delta and acceleration coefficients are not compensated as the tested noise is stationary. The whole process is illustrated in Fig. 1. The final results of the SCRS performance with PMC adapted silence model are presented in Tab. 7.

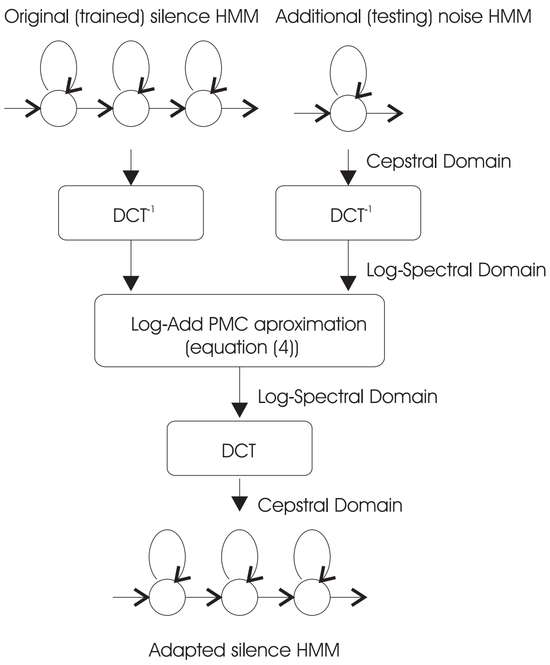

Fig. 1. Log-add PMC silence model adaptation process.

3.2.4 Discussion

The following conclusions can be deduced from the analysis of silence model influence to the SCRS robustness against the additive noise. The SCRS performance is very dependent on silence model parameters even with the relatively high *SNR* assumption (*SNR* > 25 dB). Incorrect silence model setting can ruin all the SCRS performance. It is demonstrated in Fig. 2, where the spectrograms of one selected record are depicted together with the recognized commands alignment (see vertical lines bordering the commands). Three alternatives are presented. The first spectrogram shows the record without artificially added noise and when unmodified SCRS is used. The second spectrogram shows the same SCRS output when F_3 noise is occurred with 15dB *SNR*. The third one represents the results when the silence model is PMC adapted for the same noise conditions. It can be observed from the second spectrogram that the SCRS with an inadequate silence model is unable to produce correct word borders. Furthermore F_3 noise spectrum is similar to the unvoiced phonemes spectra. This causes substitutions of correct commands by the one, which contain the phoneme with similar spectra to the background noise.

On the contrary the SCRS system indicates almost the same robustness against all testing noises with well dynamically adapted silence model. In our case almost independently on the F_1, F_2, F_3 or white noise the recognition score was in the mean higher than 90 % for speech signal with SNR higher than 15 dB (Tab. 6, 7). The same or better results were obtained with dynamically adapted silence model than with the silence model obtained by retraining. The interesting result is that with the use of narrowband noises (F_1, F_2, F_3) and corresponding silence model the less sharp decreasing recognition results were measured in comparison with white noise dependence. This can be observed in Tab. 6, 7. If the speech is well specified in the presence of additive noise then the recognition system is partially able to utilize the unaffected parts of speech signal.

SNR	white noise	F_1 noise	F_2 noise	F_3 noise
[dB]	Corr/Acc	Corr/Acc	Corr/Acc	Corr/Acc
40	97.9/95.2	98.1/93.8	97.9/94.6	97.5/93.4
30	97.9/96.5	98.3/93.8	98.5/94.8	97.7/95.5
25	97.7/96.3	98.3/95.7	97.7/94.4	96.3/94.4
20	95.9/94.6	97.1/95.9	94.6/91.3	92.8/91.5
15	94.0/92.5	89.4/88.2	87.0/84.5	87.4/86.3
10	82.0/78.1	72.2/67.9	72.5/68.1	77.6/76.2
5	59.4/53.8	50.3/45.8	52.2/44.7	67.1/65.6
0	30.4/27.3	30.0/26.5	35.6/27.1	54.9/52.8
-5	14.7/14.1	18.8/16.6	22.8/17.2	40.0/37.1

Tab. 6. Recognition rates for adapted silence model (Parameterization MFCC_E_D_A, tr = 16 ms, M = 32 ms, θ= 2).

SNR	white noise	F_1 noise	F_2 noise	F_3 noise
[dB]	Corr/Acc	Corr/Acc	Corr/Acc	Corr/Acc
40	98.1/96.9	97.9/96.5	97.7/96.5	98.1/96.3
30	97.7/97.3	98.1/97.3	97.9/96.7	97.9/97.9
25	97.5/97.3	98.1/97.7	96.3/95.4	96.5/96.5
20	95.5/95.2	97.3/96.9	92.3/91.3	94.6/94.6
15	91.1/90.5	94.8/94.4	84.5/83.8	91.1/91.1
10	65.2/64.6	84.7/83.8	73.3/71.4	84.5/84.5
5	19.7/19.3	68.7/66.9	56.3/54.2	71.8/71.6
0	8.1/8.1	45.5/44.1	39.3/36.2	51.5/51.5
-5	9.7/9.7	24.6/23.0	23.8/21.7	28.2/27.3

Tab. 7. Recognition rates for PMC adapted silence model (Parameterization MFCC_C0_D_A, tr = 16 ms, M = 32 milliseconds, θ= 2).

During the testing process in section 3.2.2 it was found that if not the all types of natural noises (*Noise*$_1$, *Noise*$_2$ and *Noise*$_3$) were included into the silence model then the recognition results were significantly worse. This was caused by poor locating of speech presence in noisy signal. This confirms the fact that the originally trained silence model is able to absorb the specific natural noises present in telephone communication.

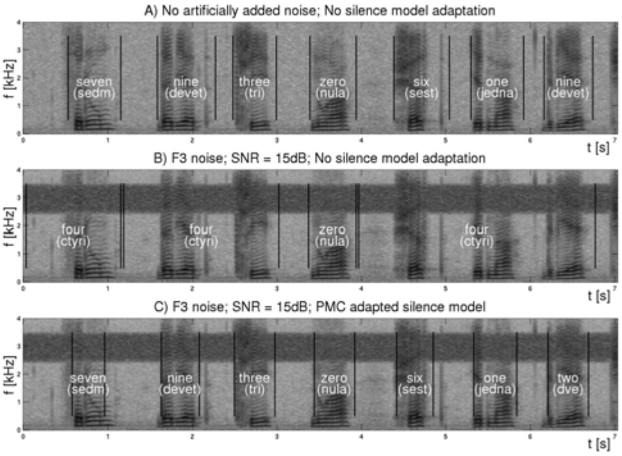

Fig. 2. Demonstration of silence model setting influence to SCRS performance.

The best results in most cases were obtained when the PMC adapted silence model was utilized. This solution unifies the advantage of the silence model trained on natural speech communication noises (*Noise*$_1$, *Noise*$_2$, *Noise*$_3$) and good adaptation on testing noises. SCRS adjusted in this way is able to operate well both in the natural communication noise and the extra stationary noise.

3.3 The Grammar Setting Influence

The SCRS has a very simple grammar definition. All possible commands are connected in parallel. The presumption is the same occurrence probability for all commands. *Grammar scale factor* and *word insertion penalty* values [15] were set to 5.0 and 0.0 respectively in all experiments. This simple grammar can be defined at least in two different ways with respect of silence model. The grammar a) (Fig. 3) where each command has to be separated by silence was used in all previous simulations. This means the need of silence model passing after every recognized command. In the alternative grammar definition b) (Fig. 3) there is no need (just the possibility) of silence model passing after the every recognized command. The results for this type of grammar are introduced in Tab. 8. These results were measured for the same set of HMM as for the results in Tab. 4. The simulations show better results for grammar a) with need of silence model passing after the each command (Tab. 4). It is probably caused by better determination of speech signal occurrence in additive noise by this type of grammar.

4. Discussion and Results

The influence of *M*, *tr* and θ parameters on the SCRS performance in noisy environment was analyzed. The correlation between appropriate setting of these parameters and the correct detection of pauses in input signal was found. The importance of the energy coefficient and its possible replacement by the C$_0$ coefficient was evaluated.

The silence model setting influence on the SCRS performance in additive noise was investigated. The improvement of SCRS robustness against the additive noise with (*SNR* > 15 dB) by the adapted silence model was observed. As expected, the PMC method was confirmed as very efficient for the silence model adaptation to a stationary noise.

Two possibilities of grammar construction with respect of silence model were analyzed.

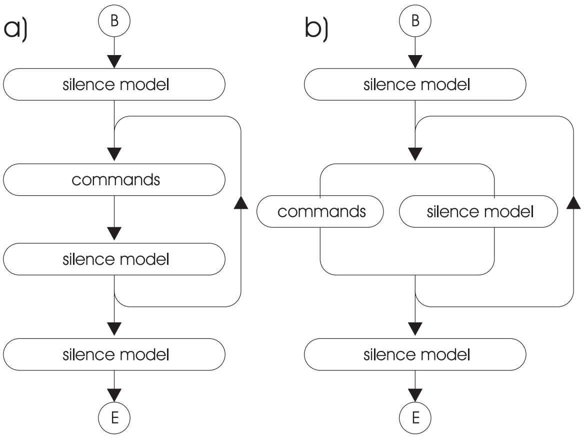

Fig. 3. Grammar definitions.

SNR	white noise	F_1 noise	F_2 noise	F_3 noise
[dB]	Corr/Acc	Corr/Acc	Corr/Acc	Corr/Acc
40	98.1/94.2	97.9/93.4	97.7/93.0	98.1/91.9
30	97.7/94.4	98.3/87.4	98.1/89.4	95.5/78.0
25	96.7/94.2	97.7/72.9	96.1/79.5	90.9/64.6
20	95.0/91.5	94.6/54.4	91.3/66.2	85.1/46.4
15	90.7/84.3	88.4/33.3	77.2/44.5	77.8/33.3
10	76.0/65.4	75.6/12.8	62.5/26.7	71.0/25.3
5	45.5/35.0	49.7/-1.4	44.7/19.0	61.1/22.6
0	14.3/13.7	29.6/1.7	29.8/20.3	52.8/19.9

Tab. 8. SCRS robustness results for grammar b).

5. Next Research

The integration of methods for additive noise influence reduction into SCRS will follow this work. From the analyses presented in this paper follows that the methods for additive noise influence reduction should be tested together with silence model adaptation. The appropriate noise parameters estimation methods with respect to telephone communication will be investigated.

Acknowledgements

This work was supported by COST 278, MSM 212300014 Research in information and communication technologies, Ministry of Education, Czech Republic and by grant GAČR 102/02/0124: Voice Technologies for Support of Information Society.

References

[1] BELLEGARDA, J. R. Statistical techniques for robust ASR: review and perspectives. *Eurospeech'97,* 1997, p. 33 - 36.

[2] MILNER, B. P., VASEGHI, S. V. Comparison of some noise-compensation methods for speech recognition in adverse environments. *IEEE Proc.-Vis. Image Signal Processing,* 1994.

[3] HERMANSKY, H., MORGAN, N. RASTA processing of speech. *IEEE Trans. Speech Audio Processing,* 1994, vol. 2, pp. 578-589.

[4] GRÉZL, F. Effect of normalization on TRAP based systems in ASR. In *Radioelektronika*, conference proceedings, 2003.

[5] KREISINGER, T., POLLÁK, P., SOVKA, P., UHLÍŘ, J. Experimental study of speech recognition in noisy environments. *Signal Analysis and Prediction,* 1998, Birkhäuser, Boston.

[6] ČERNOCKÝ, J., POLLÁK, P., HANŽL, V. Czech recordings and annotations on CD's - Documentation on the Czech Database and Database Access. *Research Report*, 2000, Prague, CTU, ED2.3.2.

[7] GALES, M. J. F., YOUNG, S. J. HMM recognition in noise using parallel model combination. *Eurospeech'93*, 1993, pp. 837-840.

[8] GALES, M. J. F., YOUNG, S. J. The application of parallel model combination to a large vocabulary dictation task. *Eurospeech'95*, 1995, pp. 1983-1986.

[9] HUNG, J., SHEN, J., LEE, L. New approaches for domain transformation and parameter combination for improved accuracy in parallel model combination (PMC) techniques. *IEEE Trans. on Speech and Audio Processing*, 2001, vol. 9, pp. 842-855.

[10] HILGER, F., NEY, H. Noise level normalization and reference adaptation for robust speech recognition. *Proc. ASR-2000*, 2000, pp. 64-68.

[11] VETH, J., MAUUARY, L., NOE, B., WET, F., SIENEL, J., BOVES, L., JOUVET, D. Feature vector selection to improve ASR robustness in noisy conditions. *Eurospeech'01*, 2001.

[12] ACCAINO, S., TSIPORKOVA, E., HAMME, H. Modelling of extra events for telephony. In workshop proceedings *Voice operated telecom services: Do they have a bright future?*. Ghent, Belgium, 2000, pp. 75-78.

[13] EALEY, D., KELLEHER, H., PEARCE, D. Harmonic tunnelling: tracking non-stationary noises during speech. *Eurospeech'01,* 2001.

[14] ANDRASSY, B., VLAJ, D., BEAUGEANT, CH. Recognition performance of the Siemens front-end with and without frame dropping on the Aurora 2 database, *Eurospeech'01*, 2001.

[15] YOUNG, S. *The HTK Book (for HTK Version 3.1)*, Cambridge University Engineering Department, 2001.

RADIOENGINEERING REVIEWERS

April 2004, Volume 13, Number 1