Contents

32

Experimental Study of Speech Recognition in Noisy Environments

Tomáš Kreisinger[1]
Pavel Sovka[1]
Petr Pollák[1]
Jan Uhlíř[1]

ABSTRACT
Achieving reliable performance in a speech recognizer for car telephone applications has been studied intensively for more than a decade. This paper addresses the effects of mismatched conditions and their minimization with respect to the performance of speaker-independent isolated-word recognition in a car-noise environment without considering the Lombard effect. This study is primarily intended to evaluate the dependence of the recognition rate on the signal-to-noise ratio (SNR) of an input signal either without any noise-compensation method or with a noise-compensation or noise-adaptive method and especially to find the appropriate conditions so that an isolated word recognizer can be used in a real car-noise environment. When hidden Markov models (HMMs) are trained on noisy speech with a SNR of 10dB, it is possible to recognize noisy speech with a SNR in the interval from 40 dB to 5 dB with a recognition rate better than 93%. If modified spectral subtraction is used and models are trained on the enhanced speech, the SNR interval increases to 0 dB. If the parallel model combination (PMC) technique is used, there is no need to train models on noisy or enhanced speech. The model adaptation enables recognizing noisy speech with any SNR from 40 to -10 dB with a recognition rate greater than 73% (for a SNR from 40 to 5 dB, the recognition rate is above 93%). In this respect PMC offers great flexibility with better recognition rates than other noise-compensation techniques.

32.1 Introduction

Isolated word recognition is applicable in many systems. We assume a voice control of dialing for mobile telephony. That is why the vocabulary used is limited to 12 words.

The utterances recorded in a stopped car are used as "clean" speech . It means that this speech set is not influenced by the variability of pronunciation caused by stress, tiredness, noise, and other driving conditions (Lombard effect). Speech and

[1] Czech Technical University, Faculty of Electrical Engineering, Technická 2, 166 27 Praha 6, Czech Republic, E-mails: {kreising, sovka, pollak, uhlir}@feld.cvut.cz

noise are recorded under the same acoustic conditions which allows generating noisy speech with various SNRs and therefore the effects of mismatched conditions[2] on the recognition rate can be analyzed.
The dependence of the recognition rate on the SNR of the test signal is studied under the following conditions:

- the training phase is done on a clean speech signal and no compensation technique is used (this case represents the lower performance of a speech recognizer);
- the training phase is done on a noisy speech signal (if SNR used for training and testing is the same, then this case represents the upper performance of a speech recognizer - in figures marked as "matched");
- some noise-compensation or noise-adaptive technique is used.

The goals are as follows:

- to verify the type of signal on which the recognizer should be trained: clean speech, noisy speech, or processed (enhanced) speech; which SNR should be used?
- to test how some chosen technique can adapt the whole recognition system to all possible input SNRs.

To answer these questions, the dependence of the recognition rate on the input SNR is evaluated. As a criterion, the SNR for which the recognition rate does not fall below a certain threshold (for our study, it is set to 93%) is used.

Some spectral subtraction methods were chosen as noise compensation techniques. Wiener filtration (the best noise-compensation technique [11]) was not used because it can be considered a special case of the parallel model combination technique (PMC). Noise-adaptive methods are represented only by PMC in this study which is one of the best methods for noise-resistant HMM recognizers [4], [11].

32.2 HMM Recognizer

32.2.1 Recognizer Construction

The recognition system is a continuous density HMM with eight states and without skips (simple left-right model). For each state in the model, the distribution is represented by an unimodal Gaussian density (no mixtures are used). No streams are used. The feature vector is composed of eight static coefficients, eight delta (or dynamic) coefficients, one energy coefficient, and one delta energy coefficient. The syntax diagram for connected word recognition contains 12 word models connected in parallel (for 12 words used in one utterance) and one model for a silence (pauses) or background noises. The weighting factors (p and s) for the transition from a word model to the silence model were varied to find their optimum for our database.

[2] Mismatched conditions mean that the different noisy conditions are used for the training phase and for the recognition.

32.2.2 Database Used and Processing Parameters

Following are the test conditions: sampling frequency 8 kHz; 16 bits per sample; frame length 20 ms; overlap 10 ms; Hamming window; number of speakers 14; number of utterances 212; number of words in one utterance eight on average; number of utterances for the training part of the database 123; number of utterances in the test part of the database 89.

Noises were recorded under various driving conditions in three types of cars. Noises are divided into three groups: stationary, nonstationary (relatively slow changes), and highly nonstationary noises (very fast changes).

Signals from speech and noise databases were mixed according to the specific SNR and used in training and testing (recognition) phases.

32.2.3 Parametrization

The parametrization used is as follows (the notation according to HTK):

a) the type of parametrization: MEL spectra = MELSPEC, MEL cepstra = MFCC, LPC spectra = LPC, LPC cepstra = LPCEPSTRA. The order for cepstral parameters is set to 20. Final cepstral vectors are then truncated to eight coefficients. The order for spectra is set to eight;

b) the type of feature vector for par=MFCC or LPCEPSTRA:
$[c_1 \ldots c_8 \, \Delta c_1 \ldots \Delta c_8 \, E \, \Delta E]$ stands for par_D_E, where E = energy.
$[c_1 \ldots c_8 \, \Delta c_1 \ldots \Delta c_8]$ = par_D, $[c_1 \ldots c_8 \, E]$ (par_E), $[c_1 \ldots c_8]$ = par.

32.2.4 Recognizer Behavior in Noisy Environments

The influence of parametrization is illustrated in Figure 32.1:

- cepstra give higher recognition rates than spectra because of better orthogonalization properties of cepstra;
- for higher SNR the LPC cepstrum is better than the MEL cepstrum because the LPC spectral modeling requires generally less data than DFT spectral modeling. On the other hand LPC is more sensitive to noise disturbances;
- the MEL cepstrum (MFCC) gives the best noise immunity and the MEL spectrum the worst;
- speech recognition rates for matched conditions represent the best performance the recognizer can achieve.

The influence of various types and numbers of parameters on the MFCC feature vector is shown in Figure 32.2. The best performance can be achieved by using static and dynamic coefficients (MFCC_D_E). The better noise immunity excludes energy and delta energy (cases MFCC_D or MFCC). This expected behavior can be explained by using spectra in the linear domain. The changes of the speech spectrum for various SNRs are greater when the energies of speech and noise are used. If the energy is omitted, then spectral changes are less which means that the models trained without using the energy are less sensitive to the disturbing noise.

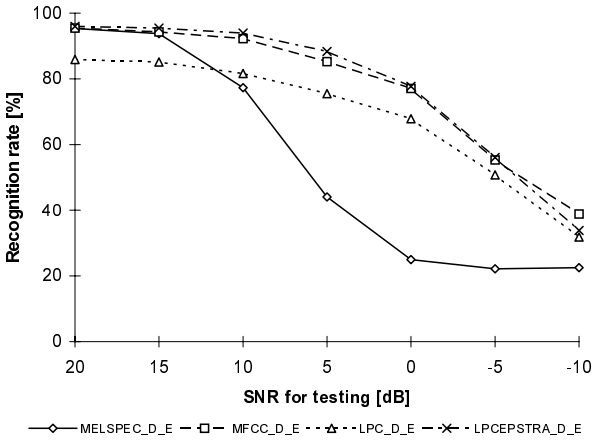


FIGURE 32.1. Recognition rates for various types of parametrization.

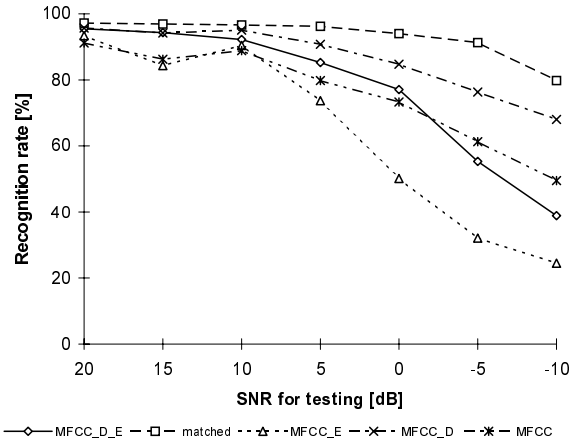


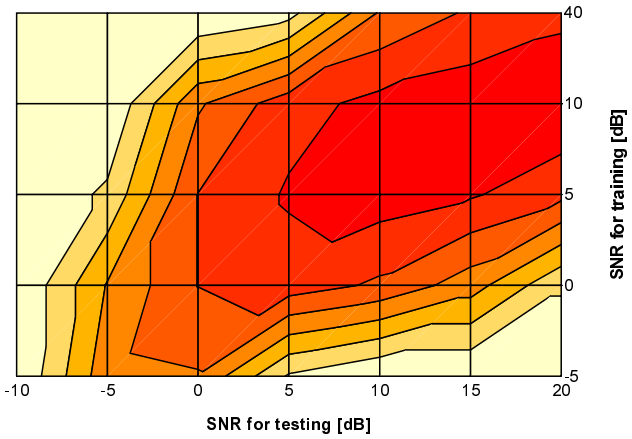
FIGURE 32.2. Recognition rates for various types of MFCC feature vector.

On the other hand, models trained this way do not represent the proper spectra of noisy speech. Therefore the recognition rates are less than when the energy is used.

For further experimentation we chose MFCC_D_E parametrization. The results are shown in Figure 32.3 which contains the table and its map for the training/testing procedure on various SNRs. This figure suggests that if no noise compensation is used then the best choice is to train models on noisy speech as follows (recognition rate should not fall below 93%!):

- training with a SNR of about 10 dB if the expected SNR of the recognition is above 5 dB ;
- training with a SNR of about 0 dB if the expected SNR of the recognition is in the range from 10 to 0 dB.

That means that we must use two different sets of HMM for SNRs above 0 dB. These conclusions are valid if the model for pauses is substituted by the model for background noise (not only the models for words but also the model for pauses are modified by training on noisy speech).



SNR tr. [dB]	SNR Testing [dB]						
	20	15	10	5	0	-5	-10
40	95.47	94.30	92.23	85.23	77.07	55.31	38.86
10	97.80	97.28	96.63	95.21	91.71	78.11	55.57
5	94.56	96.24	97.02	96.24	94.04	86.40	73.83
0	86.27	90.93	93.65	95.08	94.04	90.16	78.63
-5	45.21	62.31	77.98	85.75	91.84	91.32	79.79

FIGURE 32.3. Map of recognition rates for training HMM MFCC on noisy speech.

The diagonal of this table (or this map) represents matched conditions (the models are trained and tested under similar noise conditions with the same SNR). The first row represents the case when the training is done on clean speech and testing on noisy speech. The first case represents the upper performance of a system and the latter case represents the lower performance.

For matched conditions the influence of the various types of training procedures on the recognition rate was also tested (see Figure 32.4). First we trained the HMM on clean speech. Then we retrained these models on noisy speech keeping either variances or means constant which means that we adapted either variances (marked "const-var") or means ("const-means") to the noisy speech signal. It can be seen that omitting the variances adaptation causes a smaller decrease in the recognition rate than omitting the means adaptation. But both cases differ little from the full training (marked as "matched"). Also the differences between normal training and fixed-variance ("fix-var") training seem to be small for our database. That is why we decided to use the PMC technique for means only and not for variances (see section 32.4).

32.3 Noise Compensation

For the purposes of this study we have focused on spectral subtraction techniques only. Various spectral subtraction techniques have been frequently used for noisy speech preprocessing [1], [2], [6], [7], [8], [9] and they also have proved the efficiency for the speech recognition.

Spectral subtraction algorithms cause speech distortion as the result of noise spectrum variation in time. To evaluate how this speech distortion influences the recogni-

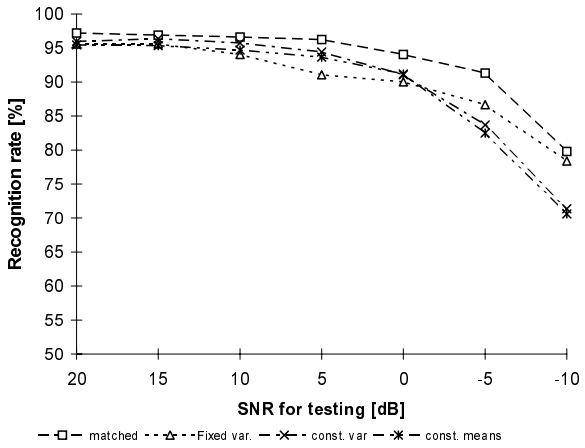


FIGURE 32.4. Various types of training for matched conditions

tion rate, raw spectral subtraction without any additional postprocessing was used. We tested spectral subtraction with half-wave rectification (acronym "so1hw") and twice repeated spectral subtraction with full-wave rectification (acronym "so1ff") [13]. The repetition of spectral subtraction allows comparable noise suppression and speech distortion in the case of half-wave rectification. Since the influence of speech distortion is studied, both algorithms use voice-activity detection according to manually created labels of signal database. The results achieved by simulations are as follows:

- HMM trained on clean speech:
  - for SNRs above 0 dB the recognition rates fall more quickly for the half-wave rectification than for full-wave because of the higher speech distortion (the difference is about 1%);
  - for SNRs below 0 dB half-wave rectification gives better results. The difference increases linearly from 1% to 7% as the SNR decreases from -5% to -10%;
- if the speech distortion is compensated for training on enhanced speech then the recognition rate is increased. This can be seen from the rows of Figure 32.5 showing the table and its map for full-wave rectification "so1ff";
- the comparison of Figures (tables) 32.3 and 32.5 reveals that spectral subtraction gives better results for lower SNRs (last three columns of both tables).

To suppress the speech distortion caused by noise spectrum variation in time and VAD failures, methods other than spectral subtraction must be used. These methods can track noise spectrum variations even during speech activity. Some examples of these methods are RASTA [7], and recently suggested Martin's [10] and Doblinger's [5] (acronyms "martin" and "dobl"). These methods track spectral minima in frequency subbands to estimate the background noise spectrum. Both methods have very good performance in nonstationary environments. The Martin method requires more computational effort than Doblinger.

a)	-5 dB	0 dB	b)	-5 dB	0 dB	c)	-5 dB	0 dB
exten	6.72	4.34	exten	7.45	4.70	exten	7.38	4.70
so1ff	7.95	8.52	so1ff	5.72	6.25	so1ff	4.84	2.69
dobl	5.20	3.58	dobl	6.72	4.10	dobl	6.85	3.25
martin	6.28	4.10	martin	7.58	4.47	martin	7.58	4.44

TABLE 32.1. SSNRE for Contamined Noisy Speech : a) stationary, b) nonstationary, c) highly nonstationary noises with SNR 0 and -5 dB.

We tried to use another approach combining spectral subtraction with an adaptive Wiener filter [15] (called extended spectral subtraction, acronym "exten"). This method uses only one parameter p controlling the smoothing of the background noise spectrum (the optimal value of p seems to be 0.95). The average noise suppression is 6 dB independent of the stationarity of noises; see Table 32.1. As the criterion the segmental signal-to-noise ratio enhancement (SSNRE [dB]) was used:

$$SSNRE = \frac{1}{L}\sum_{i=1}^{L} 10 \log \frac{P_N[i]}{P_{NR}[i]} \tag{32.1}$$

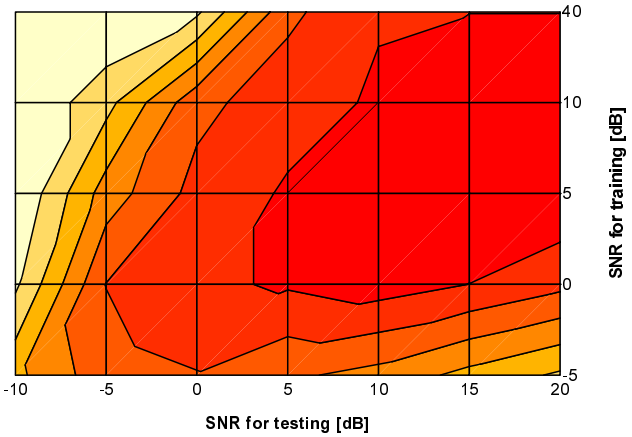
where P_N and P_{NR} are the powers of noise and residual noise respectively.

Tables 32.1(a) - 32.1(c) show the performance of the methods discussed for various types of noise from the standpoint of stationarity. By comparing the results in these three tables, it is possible to conclude that full-wave rectified spectral subtraction is better than Martin's or Doblinger's methods or extended spectral subtraction for nearly stationary noises. But for highly nonstationary noises, full-wave rectified spectral subtraction fails whereas the other three methods give noise suppression similar to the preceding case. Figures 32.5 and 32.6 illustrate the differences between spectral subtraction and extended spectral subtraction[3]. The first rows of these figures are shown in Figure 32.7. The differences in recognition rates are smaller than one should expect on the basis of the SSNRE results in Table 32.1 because all three types of noise were used simultaneously[4]. The differences in speech rates become evident especially for lower SNR. Martin's method shows the best noise immunity. The worse result of full-wave rectified spectral subtraction compared to the result of half-wave rectified spectral subtraction for lower SNR is caused by the spectral subtraction repetition which generates echoes in the processed speech. It follows from Figures 32.5 and 32.6 that if the recognition rates are to be greater than 93%, the models should be trained on enhanced speech generated from noisy speech with a SNR of about 10 dB. This approach ensures the proper recognition for noisy speech with a SNR above 0 dB. Training on enhanced speech forces the models to be less sensitive to speech distortion caused by the filtration of noisy speech. Contrary to the case of speech recognition without any noise-compensation preprocessor (see discussion for Figure 32.3), we need not use two different sets of models.

Another very often used noise-compensation method is Wiener filtration. We did not use this method because it can be considered a special case of the PMC (see next chapter).

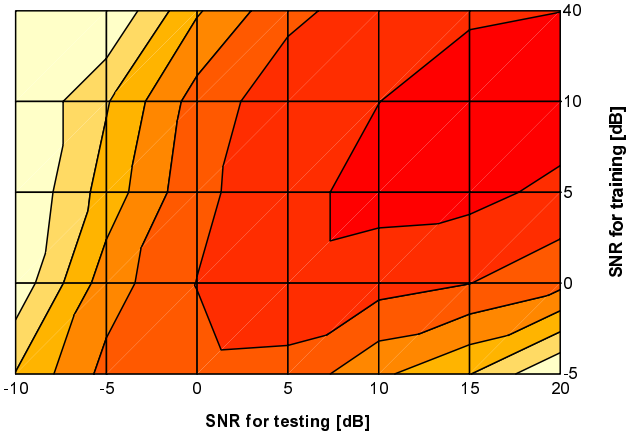
[3]Martin's method gives slightly better results than extended spectral subtraction and Doblinger's method gives worse resultes.

[4]For highly nonstationary noises only, the differences in the recognition rates are greater.



SNR tr. [dB]	SNR Testing [dB]						
	20	15	10	5	0	-5	-10
40	95.98	95.98	95.85	93.52	85.62	76.30	56.35
10	97.41	96.89	96.24	95.21	93.39	87.31	72.93
5	97.67	97.41	97.02	96.24	94.69	90.93	80.44
0	94.56	95.98	97.02	96.24	95.60	94.04	85.62
-5	87.69	89.38	91.32	92.36	93.91	93.26	89.51

FIGURE 32.5. Recognition rates for twice repeated full-wave rectified spectral subtraction.



SNR tr. [dB]	SNR Testing [dB]						
	20	15	10	5	0	-5	-10
40	95.98	95.73	94.95	93.52	89.77	82.90	68.78
10	97.28	97.02	95.98	95.21	92.88	87.82	73.58
5	95.47	96.63	96.76	95.34	93.52	88.86	80.83
0	92.62	94.04	94.82	95.34	94.04	91.06	84.59
-5	83.94	88.08	90.41	93.39	93.52	92.62	88.08

FIGURE 32.6. Recognition rates for extended spectral subtraction.

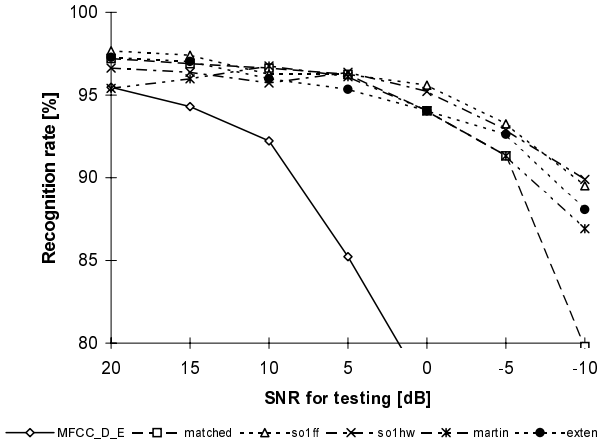


FIGURE 32.7. Comparison of noise-compensation methods.

32.4 Noise-Adaptive Methods

The problem analyzed in the foregoing section is that the important parts of speech may be removed from the signal. This is caused by filtering noisy speech by a noise-compensation method. Noise-adaptive methods are based on the adaptation of HMM parameters (trained on clean speech) to the noise whereas noisy speech is not filtered and therefore not distorted. We decided to use the noniterative PMC method [12] which modifies state probabilities and not transition ones. Moreover the frame-state alignment is not changed by this method. Since training on noisy speech affects all these three characteristics, the matched conditions represent the upper performance of the speech recognizer.

As mentioned in Section 32.2, we chose the Log Add approximation which adapts only means of HMM and not variances. This is possible because of the small database. When only static parameters are used, the PMC Log Add approximation gives results similar to Wiener filtering results [11]. Wiener filtering can be seen as adapting the static means of HMM models. This approach improves results, but there is an even more efficient PMC Log Normal approximation which can update the means and also the variances of HMM models.

The PMC Log Add approximation for static parameters is described by

$$\hat{\mu} = \mu + g\tilde{\mu} \tag{32.2}$$

where g is a weighting factor and $\tilde{\mu}$, μ, and $\hat{\mu}$ are the means of the models for the noise, clean speech, and the adapted model for noisy speech, respectively. The influence of g on the PMC behaviour was analyzed in [4]. We verified that the best results can be achieved if this factor is set according to the expected SNR during recognition. Contrary to [4] and [11], we used g as the multiplicative factor for the noise means only. We studied two possibilities: to transform $cepstra \longrightarrow log_{spectra} \longrightarrow lin_{spectra}$ and back considering the energy and without the energy. This leads to two different ways of computing the weighting factor g, but results are the same. The first way uses normalized spectra without any information about the signal or noise energy. The latter deals with spectra containing the information about the signal/noise energy.

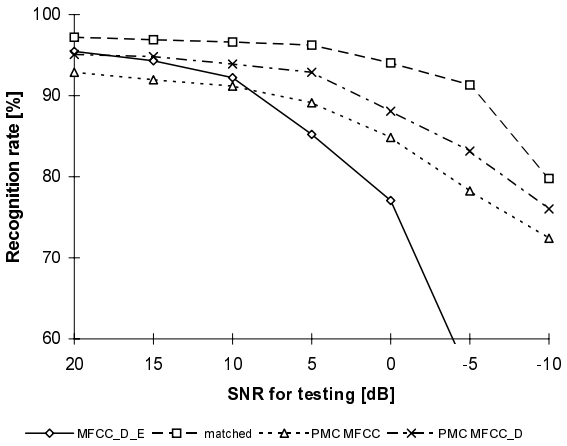


FIGURE 32.8. Performance of PMC for static and dynamic model parameters

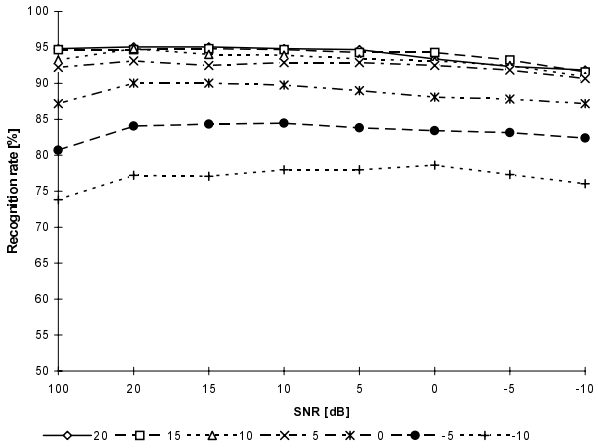


FIGURE 32.9. Influence of PMC weight g for static and dynamic coefficients.

In agreement with [3] and [4], to estimate $\tilde{\mu}$ we used approximately one second of the noisy signal preceding the word being recognized. To reliably detect nonspeech segments, a robust real-time VAD [14] should be used.

Results achieved by PMC applied to static only (PMC MFCC) or static and dynamic (PMC MFCC_D) cepstral parameters can be seen in Figure 32.8. If PMC uses both the static and dynamic coefficients simultaneously, the recognition rates are better when only static coefficients are used.

The influence of weighting in the PMC is summarized in Figure 32.9 (the case with normalized spectra) where SNR[dB] represents the expected SNR of noisy speech which equals $20 \log_{10} g$. Each type of line is generated for the given test SNR varying from 20 to -10 dB (see the legend under the figure). The maxima of the recognition rates corresponding to the proper value of g are rather flat. Similar results are also valid for static parameters. It follows from this fact that finding the proper value g (given by the expected SNR) is not a critical problem.

32.5 Conclusions

Models for connected word recognition also contain a model for pauses. If the SNR increases, this pause model becomes inadequate to describe background noise between words, and that is why the recognition rate immediately falls below 95% even when the SNR is relatively high (10 dB). The value of the weighting (penalty) factors p and s must be carefully balanced in this case. This disadvantage can be partially suppressed by training on a noisy speech signal. Training HMM, in noisy conditions similar to those expected in the recognition, forces models to learn features of speech and also features of the noise environment. This approach depends on the SNR of the input signal. Therefore it is necessary to use a noise-compensation method as the preprocessor for noisy speech. Unfortunately, these methods may cancel or distort important parts of speech. This distortion leads to lower recognition rates. The decrease of recognition rates can be partially suppressed by training models on enhanced (distorted) speech. When this idea is used, the models still depend on the SNR although less. The best results can be achieved by using the noise-adaptive method PMC. This method adapts the model parameters leaving the speech uncanceled. Therefore this approach is independent of the SNR. For real-time implementation of PMC, it is necessary to detect nonspeech segments and the SNR. Less than one second of nonspeech signals is sufficient to estimate noise characteristics.

Acknowledgments: This study was supported within the project COST 249 "Speech recognition over telephone line" and the grant GACR 102/96/k087 "The theory and application of speech communication in Czech".

32.6 References

[1] M. Berouti, R. Schwartz, and J. Makhoul. Enhancement of speech corrupted by acoustic noise. In *Proceedings of ICASSP*, April 1979, pp. 208–211.

[2] S. F. Boll. Suppression of acoustic noise in speech using spectral subtraction. *IEEE Trans. on ASSP*, 27(2):113–120, 1979.

[3] M. K. Brendborg and B. Lindberg. Noise robust recognition using feature selective modelling. In *Report of COST 249*, Roma, Italy, 1997.

[4] M.K. Brendborg. Toward noise immune automatic speech recognition using phoneme models. Ph.d. thesis, Aalborg University-Center for PersonKommunikation, Aalborg, Denmark, 1996.

[5] G. Doblinger. Computationally efficient speech enhancement by spectral minima tracking in subbands. In *Proceedings of EUROSPEECH'95*, Madrid, Spain, September 1995, pp. 1513–1516.

[6] Y. Ephraim and D. Malah. Speech enhancement using a minimum mean-square log-spectral amplitude estimator. *IEEE Trans. on ASSP*, 33(6):443–445, 1985.

[7] H. Hermansky and N. Morgan. Rasta processing of speech. *IEEE Trans. on SAP*, 2:578–579, 1994.

[8] G. S. Kang and L. J. Fransen. Quality improvement of LPC-processed noisy speech by using spectral subtraction. *IEEE Trans. on ASSP*, 37(6):939–942, 1989.

[9] P. Lockwood and J. Boudy. Experiments with nonlinear spectral subtractor (NNS), hidden markov models, and the projection for robust speech recognition in cars. *Speech Communication*, 11:215–228, 1992.

[10] R. Martin. Spectral subtraction based on minimum statistics. In *Proceedings of EUSIPCO'94*, Edinburgh, Scotland, U.K., September 1994, pp. 1182–1185.

[11] B. P. Milner and S. V. Vaseghi. Comparison of some noise-compensation methods for speech recognition in adverse environments. *IEE Proc.-Vis. Image Signal Process.*, 141(5):280–288, 1994.

[12] M.J.F.Gales and S.J.Young. HMM recognition in noise using parallel model combination. In *Proceedings of EUROSPEECH'93*, Berlin, Germany, 1993, pp. 837–840.

[13] P. Pollák, P. Sovka, and J. Uhlíř. Noise suppression system for a car. In *Proceedings of EUROSPEECH'93*, Berlin, 1993, pp. 1073–1076.

[14] P. Sovka and P. Pollák. The study of speech/pause detectors for speech enhancements methods. In *Proceedings of EUROSPEECH'95*, Madrid, Spain, 1995, pp. 1575–1578.

[15] P. Sovka, P. Pollák, and J. Kybic. Extended spectral subtraction. In *Proceedings of EUSIPCO'96*, Trieste, Italy, 1996, pp. 963–966.