ACOUSTIC FRONT ENDS FOR SPEAKER-INDEPENDENT DIGIT RECOGNITION IN CAR ENVIRONMENTS

D. Langmann, A. Fischer, F. Wuppermann, R. Haeb-Umbach, T. Eisele

Philips GmbH Forschungslaboratorien Aachen
P.O. Box 50 01 45
D-52085 Aachen
Germany
Email: {langmann,afischer,wupper,haeb,eisele}@pfa.research.philips.com

ABSTRACT

This paper describes speaker-independent speech recognition experiments concerning acoustic front end processing on a speech database that was recorded in 3 different cars. We investigate different feature analysis approaches (mel-filter bank, mel-cepstrum, perceptually linear predictive coding) and present results with noise compensation techniques based on spectral subtraction. Although the methods employed lead to considerable error rate reduction the error analysis shows that low signal-to-noise ratios are still a problem.

1 INTRODUCTION

Automatic speech recognition in a car is a difficult problem due to the adverse acoustic environmental conditions [1], [2]. For example, the A-scored immission level of medium-class cars increases from around 55-58 dB(A) at 50 km/h to 67-70 dB(A) at 100 km/h and further to 71-75 dB(A) at 130 km/h. This results in signal-to-noise ratios below 0 dB in the worst case. The signal power of the noise is rapidly changing depending on car body, traffic situation, speed, and in-car acoustic events such as radio, wiper and passenger conversation. Another important source of degradation is the "Lombard effect", i.e. the change of the speech signal generation when produced in a noisy environment.

In order to obtain realistic data, recordings have been conducted in running cars under various acoustic environmental conditions (different speeds, window/radio on or off, etc.). While earlier car speech databases were primarily intended for investigation of speaker-dependent recognition [3], [4], [5], [6], the database used in this paper comprises 200 speakers and is suitable for speaker-independent tests in different car environments.

In first experiments, which we describe in this paper, we compare different acoustic front ends. The results give some indication on the robustness of standard feature sets, such as mel-filter bank, mel-cepstrum [7] and perceptually linear predictive cepstrum [8]. Further we experiment with explicit noise removal techniques such as spectral subtraction [9], [10], SNR normalization [11] and a combination thereof. This combination reduces the error rate by 30%, but the error rates obtained, 12% word error rate for a quite difficult digit string recognition task, still leave room for improvement.

2 THE CAR SPEECH DATABASE

A speech data collection has been conducted in 3 cars comprising a total of 102 female and 102 male speakers. The cars used are BMW 750i, VW Passat TDI and Ford Escort 16V CLX. Note that the speakers were the drivers of the cars. Each speaker spoke a set of 45 utterances including isolated digits, digit strings, spellings, location names, command words and phonetically rich sentences. The text material was designed to enable training and assessment of both isolated and continuous-speech utterances, employing whole-word or sub-word approaches. Controlled recordings in different acoustical environmental conditions were conducted, such as city and highway rides, radio on or off, side window open or closed, rain yes or no.

The acoustic signal was captured with two electret car microphones mounted on the car ceiling to the left and to the right of the driver. The transfer characteristics of the microphone spans the range of 500 to 5000 Hz. In the experiments only the right microphone signal was employed.

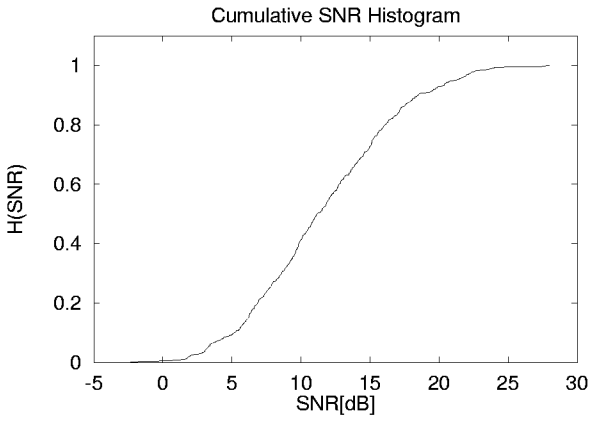

Figure 1: Total SNR distribution of the digits subcorpus of car speech data base.

In our tests we investigated the digits subset of the car speech data collection. The average SNR of this subcorpus is 11.6 dB with SNR values ranging from -2.3 dB to

25 dB (see fig. 1). For training we selected 87 female and 87 male speakers which had uttered 2068 digits in total. For the recognition tests, the recordings of the remaining 15 female and 15 male speakers were used. They spoke 773 digits. Both in training and test set, use of the 3 car types is equally distributed. There is a slight mismatch between average training and test SNR. The average SNR in training set is 12 dB and in the test set 10.8 dB. Note that training and test is carried out in noisy conditions (no mismatch in that sense). The very inhomogeneous data in training and test has to be coped with.

3 ACOUSTIC PREPROCESSING

3.1 Sampling Rate

First we investigated the influence of the sampling rate on the recognition performance. It is known from hearing experiments with systematically high-pass and low-pass filtered speech that the frequency range of 300 to 5000 Hz contains the perceptually most important frequencies. Indeed, our experiments confirmed that a sampling rate of 8 kHz (signal bandwidth 4 kHz), as is standard for telephone speech, leads to a degradation of the recognition performance by 5-10%, compared to a sampling rate of 11.025 kHz (signal bandwidth 5.5 kHz). In the following we therefore used the higher sampling rate.

3.2 Short-Term Feature Analysis

The speech signal is sampled at 11025 Hz, preemphasized and blocked into 32 ms frames by a Hamming window. Then a 512-point FFT is performed with 16-ms shift. Three sets of feature vectors have been compared:

- MTFB: Mel-spaced triangular filter bank. The resulting power spectrum is convolved with a triangular filter kernel and then sampled at 16 frequencies arranged roughly linearly on a mel-frequency scale. Logarithm is then applied to the filterbank outputs [7].
- MFCC: Mel-frequency cepstral coefficients. A discrete cosine transform is applied to the MTFB coefficients, and the first 12 cepstral coefficients are retained.
- PLP: Perceptually linear predictive coding. PLP is an approximation of auditory-like spectrum by autoregressive all-pole modeling. It takes into account critical-band spectral resolution, 40-dB equal-loudness curve and the Steven's intensity-loudness power law [8]. We utilized the first 12 PLP-derived cepstral coefficients here.

Each of the three different feature vector types was subsequently subjected to the following operations:

- High-pass filtering. Each feature vector component trajectory was filtered by a first-order high-pass filter in order to reduce the influence of a changing acoustic environment.
- Augmentation of the feature vector by linear regression coefficients. In our experiments it turned out that delta coefficients computed by linear regression performed consistently better than simple first-order time differences. The regression coefficients were computed over a window of 64 ms.

Each resulting feature vector consists of 24 components.

3.3 Spectral Subtraction

Spectral subtraction enhances speech signals through the subtraction of an estimated noise spectrum [9]. This increases the signal-to-noise ratio with the possible side-effect of introducing so-called musical noise through residual peaks in the spectral floor.

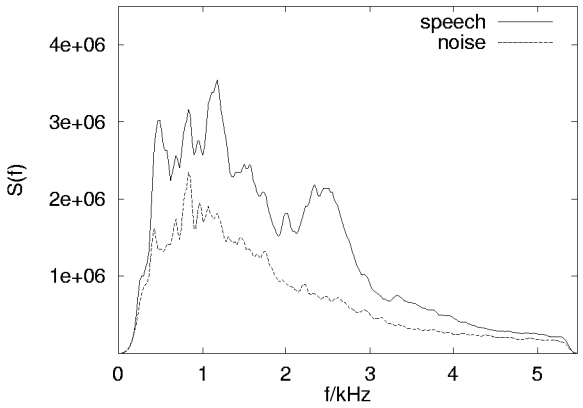

Figure 2: Spectral energy distributions of the digits subcorpus of car speech data base.

Let $S(f,t)$ denote the speech spectrum corrupted by additive noise and $N(f,t)$ be an estimate of the noise spectrum obtained form noise-only periods of the recorded signal. Note the dependence on the time t, which illustrates the time-varying nature of the spectra. Figure 2 shows the averaged spectral energy distributions separated for speech and noise. An estimate of the uncorrupted speech signal $X(f,t)$ is obtained by subtracting the noise spectrum estimate from the incoming signal:

$$\hat{X}(f,t) = \max(S(f,t) - aN(f,t)\,; bN(f,t))$$

The subtraction operation is slightly modified by an overestimation factor a for the noise spectrum and by applying a bottom clip to avoid small and negative values in the subsequent logarithm operation.

More sophisticated schemes have been introduced in order to suppress musical noise and other deficiencies. One major extension is the choice of a frequency- and/or time-variant overestimation factor $a(f,t)$ that is determined from the current signal and noise condition [10]:

$$\hat{X}(f,t) = \max(S(f,t) - a(f,t)N(f,t)\,; bN(f,t))$$

The resulting spectral subtraction method is called nonlinear spectral subtraction (NSS).

3.4 SNR Normalization

In order to make the corpus more homogenous with respect to the SNR the technique of SNR normalization [11] is used. The linear filter bank outputs $X_k(t)$, where k denotes the filter bank index, are masked with a masking value $M_k(t)$:

$$Z_k(t) = X_k(t) + M_k(t)$$

The masking value is computed as a function of the instantaneous SNR. The instantaneous SNR in turn is computed as the ratio between smoothed filter bank output signals considered as speech, $\overline{S}_k(t)$, and those considered as noise, $\overline{N}_k(t)$:

$$\mathrm{SNR}_k(t) = \overline{S}_k(t)/\overline{N}_k(t)$$

$\overline{S}_k(t)$ and $\overline{N}_k(t)$ are obtained from low-pass filtering the filter bank output signal of frames classified as speech and noise, respectively. For example in case of $Z_k(t)$ being classified as speech:

$$\overline{S}_k(t) = \alpha\overline{S}_k(t-1) + (1-\alpha)Z_k(t)$$

with some appropriate filter factor α. If the actual $\mathrm{SNR}_k(t)$ is larger than a target SNR, the masking constant is increased; if smaller, the masking constant is decreased. Thus the target SNR is tracked.

A promising approach is the combination of spectral subtraction and SNR normalization. In the case of a very noisy environment the measured SNR can drop below the target SNR even if the masking offset has been set to zero. Then no SNR normalization is possible. If the speech is enhanced by spectral subtraction, these high noise regions disappear and the subsequent SNR normalization can be more effective.

4 EXPERIMENTS

4.1 The Recognition Framework

In our experiments, we employ the Philips continuous-speech recognition framework [12]. It is based on statistical modeling of speech by left-to-right Hidden Markov Models (HMM) with Laplacian mixture densities. A state-independent diagonal covariance matrix is utilized. We make use of digit whole-word models with fixed transition probabilities allowing only loop, forward, and skip transitions. The emission probabilities are trained according to the maximum likelihood principle by an iterative estimation-maximization procedure.

The speech recognition is performed by Viterbi decoding and time-synchronous one-pass search. In addition to the valid recognition vocabulary, a background model is included as a permanent rejection alternative. There are no length restrictions on the digit strings to be recognized.

4.2 Car Speech Database

Table 1 compares the word error rates of the three different signal analysis approaches described in Section 3.2. MFCC clearly performs best. The error rate for PLP is somewhat disappointing since other sites report the robustness of PLP particularly in noisy conditions. In the following we always assume MFCC feature analysis.

	MTFB	MFCC	PLP
WER [%]	19.8	17.3	18.9

Table 1: Word error rates (WER) of different acoustic front ends.

	MFCC	NSS+MFCC	NSS+SNR+ MFCC
WER [%]	17.3	14.7	12.0

Table 2: Word error rates (WER) for non-linear spectral subtraction (NSS) and the combination of NSS with SNR normalization (target SNR: 11dB).

Table 2 shows the effect of the different noise compensation techniques described in Sections 3.3 and 3.4. As can be observed, the combination of nonlinear spectral subtraction and SNR normalization is advantageous, leading to a error rate reduction by 30%.

4.3 Control Experiments on NOISEX-92

The error rates reported above are fairly high for a digit string recognition task. Therefore we ran control experiments on the NOISEX-92 database [13]. This is a small speaker-dependent database of English digits containing clean speech and speech with artificially added noise. We downsampled the data to 8 kHz and ran experiments on the data contaminated by car noise at SNRs from -6dB to +18 dB. The above recognizer was applied without any parameter adjustment on the isolated digits test set. Training was performed on the isolated digits and digit triplet noise-free training sets.

Table 3 presents the error rates and shows the dramatic improvement for low SNR values resulting from nonlinear spectral subtraction and SNR normalization. The results obtained compare well with published results for SNRs up to zero dB [11]. Therefore we concluded that the high error rates on the car database are mainly due to the nature of the data and the very small number of training digits per speaker.

SNR [dB]	-6	0	6	12	18
MFCC	99/100	94/96	27/34	0/0	0/0
NSS+SNR+ MFCC	69/83	4/13	0/0	0/0	0/0

Table 3: Speaker-dependent digit error rates (females/ males) for 100 isolated digits test set contaminated by car noise without spectral tilt (NOISEX).

5 ERROR ANALYSIS

The error rate on the car database is fairly high compared to other digit recognition tasks, e.g. over the telephone. Therefore we made an analysis to gain insight as to where the recognition problems are. It is instructive to subdivide

the recognition corpus into different SNR ranges. Table 4 shows that utterances with low SNR still pose major problems despite the aforementioned noise compensation techniques.

SNR range [dB]	-2.3 .. 10	10 .. 15	> 15
WER [%]	21.6	7.0	5.7

Table 4: Subset word error rates according to SNR ranges.

Gender	Female	Male
Mean SNR [dB]	11.1	12.2
WER [%]	15.4	8.2

Table 5: Subset word error rates according to gender.

Further we looked into the relative performance of the male and female speakers. Although the average SNR of the utterances of the female speakers was not much lower, the error rate is almost a factor of two higher, see Table 5. The subdivision of the recognition corpus according to car type (Table 6) and traffic situation (Table 7) are also quite instructive.

Car	Ford	BMW	VW
Mean SNR [dB]	9.8	11.6	13.6
WER [%]	18.0	9.2	7.9

Table 6: Subset word error rates according to car type.

Traffic	City	Highway
Mean SNR [dB]	12.5	9.7
WER [%]	8.6	18.9

Table 7: Subset word error rates according to traffic situation.

6 SUMMARY

First results of digit recognition experiments have been reported on a car speech database for speaker-independent speech recognition in different cars. We have seen that mel-frequency cepstral coefficients perform better than both a log-spectral feature vector and features obtained from PLP analysis. Nonlinear spectral subtraction in combination with SNR normalization delivers an error rate reduction of about 30%. Currently we achieve an error rate of 12%. Despite the recent progress, robustness is still a major research issue.

7 REFERENCES

1. Juang, B. H. "Speech Recognition in Adverse Environments", Computer Speech and Language 5: pp. 275 - 294, 1991.
2. Junqua, J.-C., Haton, J.P., "Robustness in Automatic Speech Recognition: Fundamentals and Applications", Kluwer, Boston, 1996.
3. Codogno, M.; Fissore, L.; Laface, P.; Venuti, G. "HMM Modelling for Voice-Activated Mobile-Radio System", CSELT Technical Reports, Vol. 19, No. 1, pp. 41 - 44, 1991.
4. Lockwood, P.; Boudy, J.; Lelievre, L.; Nicke, H. "Vocal Dialing for Radiotelephones in Cars", Proceedings of 5th Nordic Seminar on Digital Mobile Radio Communications DMR V, pp. 243 - 246, 1992.
5. Geller, D.; Haeb-Umbach, R.; Ney, H. "Improvements in Speech Recognition for Voice Dialing in the Car Environment", In: Proceedings of the ESCA Workshop "Speech Processing in Adverse Conditions", Cannes-Mandelieu (France), pg. 203-206, 1992.
6. Mokbel, Ch. E.; Chollet, G. F. A. "Automatic Word Recognition in Cars", IEEE Transactions on Speech and Audio Processing. Vol. 3, No. 5, pp. 346-356, September 1995.
7. Davis, S. B.; Mermelstein, P. "Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences", IEEE Transactions on ASSP 38: pp. 1870 - 1878, 1980.
8. Hermansky, H. "Perceptual Linear Predictive (PLP) Analysis of Speech", J. Acoust. Soc. Am. 87(4): pp. 1738-1752, April 1990.
9. Berouti, M.; Schwartz, R.; Makhoul, J. "Enhancement of Speech Corrupted by Acoustic Noise", Proceedings of ICASSP, Washington D.C., Columbia, pp. 208-211, 1979.
10. Lockwood, P.; Boudy, J. "Experiments with a Nonlinear Spectral Subtractor (NSS), Hidden Markov Models and the Projection, for Robust Speech Recognition in Cars", Speech Communication 11: pp. 215-228, 1992.
11. Claes, T.; van Compernolle, D. "SNR-Normalisation for Robust Speech Recognition", Proceedings of ICASSP, Atlanta, Georgia, pp. 331 - 334, 1996.
12. Ney, H.; Steinbiss, V.; Aubert, X.; Haeb-Umbach, R. "Progress in Large Vocabulary, Continous Speech recognition", In: Niemann, H.; de Mori, R.; Hanrieder, G. (Eds.) "Progress and Prospects of Speech Research and Technology", infix, St. Augustin, pp. 75 - 92, 1994.
13. Varga, A.; Steeneken, H. J. M.; Tomlinson, M.; Jones, J. "The NOISEX-92 Study on the Effect of Additive Noise on Automatic Speech Recognition", Booklet included in the NOISEX-92 CD-ROM Set.
14. Le Bouquin, R. "Enhancement of Noisy Speech Signals: Application to Mobile Radio Communications", Speech Communication 18: pp. 3-19, 1996.
15. Hermansky, H.; Morgan, N. "RASTA Processing of Speech", IEEE Transactions on Speech and Audio Processing. Vol. 2, No. 4, pp. 578-589, October 1994.
16. Yang, R.; Haavisto, P. "Noise Compensation for Speech Recognition in Car Noise Environments", Proceedings of ICASSP, Detroit, Michigan, Vol. 1, pp. 433-436, 1995.