SUBWORD UNIT BASED SPEECH RECOGNITION IN CAR ENVIRONMENTS

Alexander Fischer and Volker Stahl

Philips Research Labs, Aachen
Weißhausstr. 2, D-52066 Aachen, Germany
email: {afischer,vstahl}@pfa.research.philips.com

ABSTRACT

This paper presents results of speaker-independent speech recognition experiments concerning acoustic front-ends, models and their structures in car environments. The database comprises 350 speakers in 6 different cars. We investigate whole-word models, context-independent phoneme models and context-dependent within-word phoneme models. We studied task-dependent (same vocabulary context in training and test) phoneme models and present first results on task-independent (broad context in training, i.e. phonetically rich material) scenarios. The latter allows flexible vocabulary definition for applications with dynamically changing command words or new applications avoiding an expensive data collection. Acoustic preprocessing is carried out with mel-cepstrum combined with spectral subtraction and SNR normalization. The task-dependent word error rates are well below 3% for both whole-word and phoneme models. The task-independent scenarios have to be worked on further.

1. INTRODUCTION

Speech recognition in car environments has to cope with adverse acoustic conditions [1], [2]. Immission levels of medium-class cars rise from 55-58 dB(A) at 50 km/h up to 71-75 dB(A) at 130 km/h. This can lead to signal to noise ratios even below 0 dB. The main noise influences are the car body resonance, driving speed, and other acoustic sources in the car such as radio, wiper and passenger conversation. Furthermore the Lombard effect (change of speech characteristics in the presence of background noise) has to be taken into account.

In [3] we presented results concerning acoustic front-ends for speaker independent connected digit recognition on a database of 200 speakers in 3 different cars. The database has been augmented to 350 speakers in 6 different cars (section 2.1). In this paper we present results for speaker independent recognition of command words and city names in a task-dependent manner and first results for task-independent scenarios [11, 10]. Acoustic preprocessing (section 2.2) is carried out with mel-frequency cepstral coefficients (MFCC) enhanced by nonlinear spectral subtraction [4, 5] (NSS) and SNR normalization [6, 7] being the most powerful front end investigated in [3]. HMM modeling is done by whole-word and subword (context-independent monophone and context-dependent triphone) models. Results for MFCC, MFCC+NSS and MFCC+NSS+SNR preprocessing and different model structures for task-dependent subword models are given in section 2.4. Preliminary results for task-independent scenarios follow in section 2.5. An analysis of the errors over gender and cars in section 3 highlights the major problems to be dealt with in the future.

2. EXPERIMENTS

2.1. Database

For the command words and city names we use 155 speakers in 3 different lower-class cars (VW Golf, Fiat, Hyundai) of the whole database. 116 speakers are used for training, the remaining 39 speakers are the test set for the task-dependent scenarios. The command words scenario consists of 43 words related to the control of typical car functionality (radio, phone). The city names are those of the 38 biggest German ones. Figure 1 and 2 show the SNR distribution of the speech data of both scenarios for both genders separately and joined. The SNR of each utterance was measured via a log-energy histogram. The mean SNR for males and females is indicated by the vertical lines in the plots.

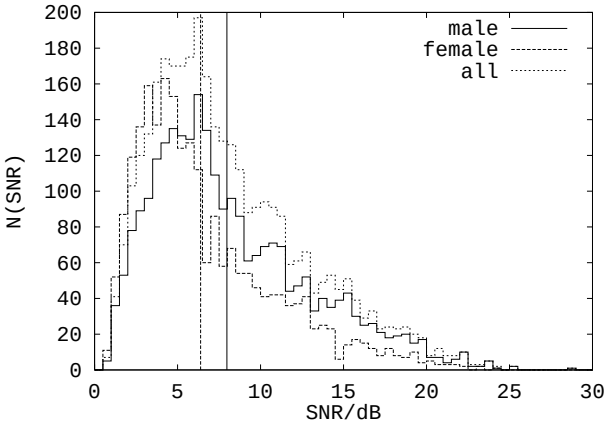

Figure 1: SNR histogram for command words

The task-independent scenarios are trained on phonetically rich material from 205 speakers in three different cars (BMW 750i, VW Passat TDI, Ford Escort). The material consists of 1813 utterances with 10978 words. Figure 3 shows the SNR distribution of the phonetically rich material.

One can clearly see the difference between the lower class cars of the command words respectively city names data (average SNR 6-7dB) and the medium to upper class cars of the phonetically rich material (average SNR 10-11dB). The average SNR of the female speakers is 1.5-2dB lower than that of the male speakers. For the task-independent investigations we have to cope with both the vocabulary context mismatch and the car mismatch.

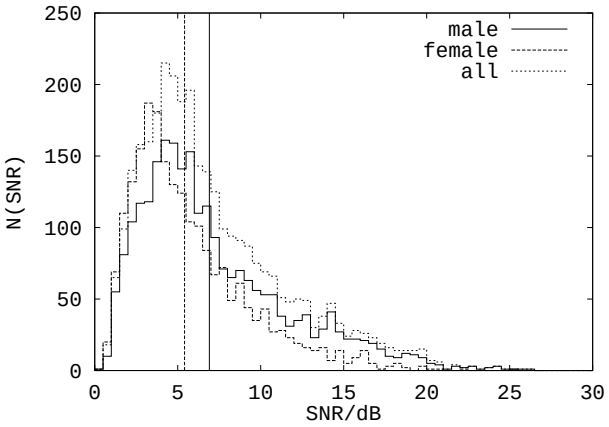

Figure 2: SNR histogram for city names

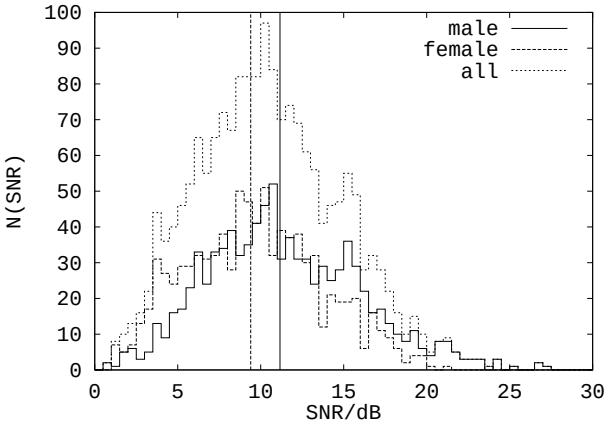

Figure 3: SNR histogram for phonetically rich sentences

2.2. Acoustic Preprocessing

The speech signal is sampled with a rate of 8kHz. The sampled signal is blocked into 32ms frames with a frame shift of 16ms. Each frame is subjected to a Hamming window followed by a 256-point FFT. The FFT power spectrum is convolved with a triangular filter kernel and sampled at 15 frequencies according to a mel-frequency scale. After the log operation on the filterbank outputs a discrete cosine transform yields 12 mel-frequency cepstral coefficients (MFCC). The influence of a changing acoustic environment is reduced by filtering each feature vector component with a first-order high-pass filter. Finally the feature vector is augmented by 8 delta coefficients that are computed by linear regression over 5 frames from the first 8 MFCCs [3]. Each resulting feature vector has 20 components.

Nonlinear spectral subtraction is used on the FFT power spectrum and SNR normalization is applied to the filterbank outputs as discussed in [3]. Spectral subtraction enhances the data (increase of SNR) whereas SNR normalization introduces noise (and thereby decreases the SNR) to those parts of the data that have a higher SNR than a given target SNR. SNR normalization has no effect on data with an SNR below the target SNR. The benefit of SNR normalization is an increased homogeneity of the data material in terms of SNR. Both methods can be promisingly combined because preliminary spectral subtraction allows SNR normalization to affect a bigger part of the data.

2.3. Recognition Framework

The experiments are carried out with the Philips continuous-speech recognition framework [9]. 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. The whole-word and phoneme models have 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 (garbage model) is included as a permanent rejection alternative. Although all command words and city names were uttered in isolated manner, there is no single word restriction during recognition.

2.4. Task-Dependent Modeling

2.4.1. Word Models

Table 1 shows the average model length (in states) and the word error rates for command words and city names when using task-dependent whole-word models with different acoustic front-ends.

	average model length	MFCC	MFCC +NSS	MFCC +NSS+SNR
app	30 ± 11	4.45	3.28	2.78
names	25 ± 9	4.62	2.80	2.60

Table 1: Average model lengths and word error rates (WER) of whole-word models for different acoustic front-ends.

The baseline MFCC front-end yields a word error rate below 5% for both scenarios. Nonlinear spectral subtraction is quite powerful on these tasks. A relative improvement of 25% for the command words and 40% for the city names can be gained. SNR normalization yields a further error rate reduction of 15% for command words and 7% for city names thus proving the combination effect of both methods. The parameters of NSS and SNR normalization have been developed on the word model scenarios and were kept fixed during the rest of the investigation.

2.4.2. Phoneme Models

We investigated context-independent (CI) phonemes (monophones) and context-dependent (CD) phonemes (triphones with monophone fallback). The model architecture is defined by the number of segments (S) and the number of identical states in each segment (Q). The 2,2 model for example has 4 states in total but the first and last 2 states share the same emission probabilities. This model architecture hence requires half the number of emission probability parameters than the 4,1 model.

Table 2 shows the word error rates of the task-dependent recognizers for different acoustic front-ends and model structures.

The baseline error rate for CI phonemes is about 8% for both command words and city names and is quite independent of the model structure. CD phonemes perform significantly better (about 5% WER) and are comparable to whole-word models. Nonlinear

model	MFCC		MFCC+NSS		MFCC+NSS +SNR	
S, Q	CI	CD	CI	CD	CI	CD
command words						
2,2	7.91	5.07	6.49	3.53	5.38	3.40
4,1	7.48	4.27	6.43	3.22	5.28	3.22
5,1	7.36	4.39	5.63	3.03	5.07	2.54
city names						
2,2	8.27	6.05	5.01	3.58	5.08	3.06
4,1	8.33	4.69	5.01	3.19	5.14	2.47
5,1	8.40	4.23	4.82	2.73	4.23	2.15

Table 2: Word error rates (in %) of context-independent (CI) and context-dependent (CD) phoneme models trained on same vocabulary context for different acoustic front-ends.

spectral subtraction and SNR normalization result in similar error rate reductions as for whole-word models. A relative reduction of 20% to 40% due to NSS and an additional 20% reduction by SNR normalization can be observed. The gain by NSS is quite consistent over the investigated scenarios whereas the SNR normalization results differ somewhat according to model structure and task (command words or city names). For the city names monophones in 2,2 and 4,1 structure even a slight degradation by SNR normalization takes place. The 5,1 model structure is clearly advantageous over the other ones and shows consistent improvement from CI to CD phonemes and for the two robust preprocessing techniques.

2.5. Task-Independent Modeling

Task-independent modeling gives the opportunity to avoid expensive data collections necessary for whole-word recognizers. In addition it allows applications with flexible vocabularies. With the same parameter settings and model structures as before CD recognizers were trained on the phonetically rich material of the car-speech database. CI training will be studied in the future. The limited amount of phonetically rich training material makes the exploitation of other larger phonetically rich databases more promising. The CI recognizers were then used to recognize the above command words and city names sets. Table 3 shows the word error rates of our first studies on task-independent scenarios for different acoustic front-ends and model structures.

model S, Q	MFCC	MFCC+NSS	MFCC+NSS +SNR
command words			
2,2	16.1	13.7	13.9
4,1	15.5	13.4	15.2
5,1	12.6	11.4	13.7
city names			
2,2	22.9	20.4	21.1
4,1	20.4	19.0	19.6
5,1	19.1	17.7	18.8

Table 3: Task-independent word error rates (in %) of context-independent (CI) phoneme models for different acoustic front-ends.

The degradation by task-independent modeling is less severe for the command words than for the city names. The command words are approximately one phoneme longer than the city names (table 1) which makes the recognition task easier. The superiority of the 5,1 model shows up for task-independent recognizers, too. Nonlinear spectral subtraction is quite effective here, too (10% to 20% relative gain). The SNR normalization part of the acoustic front-end is unable to cope with the acoustic mismach (4-5dB in average SNR, figures 1, 2 and 3) due to the different car classes in training and test. It nearly reverts the gain obtained by NSS.

CI phonemes should be able to significantly reduce the observed degradation for these task-independent scenarios [10], [11]. This is indicated by the results in table 2 representing an ideal CD phoneme set and will be studied in the future.

3. ERROR ANALYSIS

In order to gain some more insight on the main recognition problems we look at the recognition results for the 5,1 model structure of the task-independent recognizers in some more detail. A subdivision with regard to car and gender is shown in table 4.

model S, Q	MFCC	MFCC+NSS	MFCC+NSS +SNR
command words			
male	7.8	8.1	9.0
female	17.5	13.9	17.5
Fiat	15.7	13.0	15.6
Hyundai	10.2	9.2	13.4
VW Golf	12.3	11.0	11.8
city names			
male	13.7	13.3	14.0
female	24.4	21.7	24.4
Fiat	19.6	17.0	19.8
Hyundai	20.4	20.9	20.2
VW Golf	17.5	15.0	17.9

Table 4: Word error rates (in %) of context-independent (CI) 5,1 phoneme models for task-independent recognizers

Female speakers with up to twice as high error rates as male speakers pose a major problem. This can be partly explained by the lower SNR (1.5-2dB with regard to male speakers) of female speakers. A significant car dependency can not be stated. NSS mainly improves performance for female speakers (10% to 20% relative) and a similar gain can be observed over the different cars. A slight degradation of the Hyundai in the city names scenario is the only exception to that effect. The malfunction of SNR normalization in the task-independent scenarios is gender and car independent. The large global SNR mismatch of 4-5dB in average (see figures 1, 2 and 3) between the phonetically rich training in the medium/upper-class cars to the test in the lower-class cars is the main reason for this effect. Figures 4 and 5 show the error rates of the single speakers in the test sets for the baseline front-end (MFCC) and the one with spectral subtraction (MFCC+NSS). The first 19 speakers are male.

NSS clearly improves the performance for speakers with high baseline error rates ($> 20\%$) but the overall gain indicated by the horizontal lines is about 10% relative. This is not as high as for the task-dependent scenarios (up to 40% relative) and suggests a dedicated front-end for these tasks.

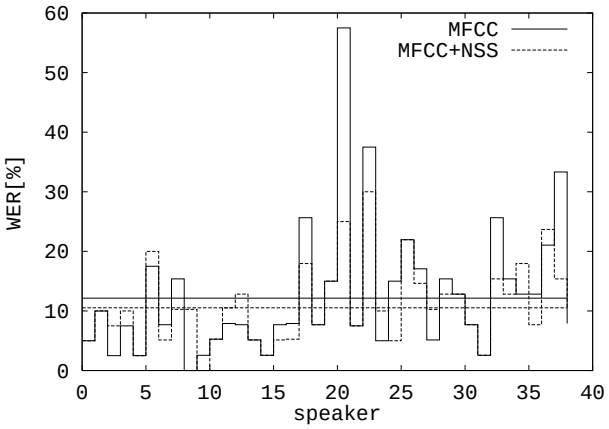

Figure 4: Speaker word error rates (WER) of command words for task-independent 5,1 monophones

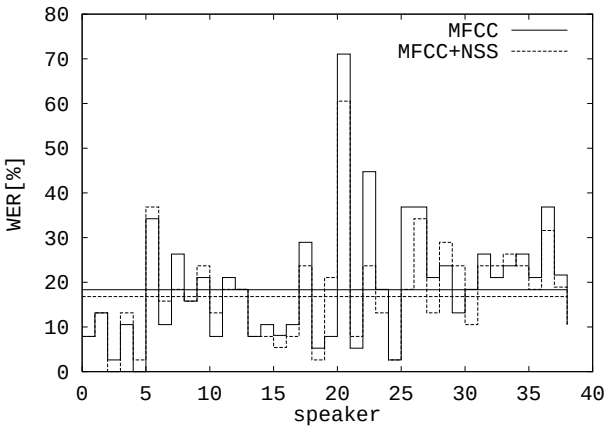

Figure 5: Speaker word error rates (WER) of city names for task-independent 5,1 monophones

4. CONCLUSION

We presented first results on command word and city name recognition in the car environment on a realistic database. Task-dependent whole-word and phoneme models with robust feature extraction using nonlinear spectral subtraction and SNR normalization yield error rates below 3%. First results on task-independent modeling using context-independent phoneme models based on phonetically rich in-car recordings show the necessity for context-dependent modeling and a dedicated front-end for such scenarios. Context-dependent training that will be studied in the future needs the exploitation of larger databases of phonetically rich material (e.g. telephone or office). The even larger acoustic mismatch then requires a dedicated front-end, too.

5. 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] Langmann, D., Fischer, A., Wuppermann, F., Haeb-Umbach, R., Eisele, T. ”Acoustics Front Ends for Speaker-Indepdendent Digit Recognition in Car Environments”, Proceedings of Eurospeech, Rhodos, Greece, pp. 2571-2574, 1997.

[4] Berouti, M., Schwartz, R., Makhoul, J. ”Enhancement of Speech Corrupted by Acoustic Noise”, Proceedings of ICASSP, Washington D.C., Columbia, pp. 208-211, 1979.

[5] Le Bouquin, R. ”Enhancement of Noisy Speech Signals: Application to Mobile Radio Communications”, Speech Communication 18: pp. 3-19, 1996.

[6] Claes, T., van Compernolle, D. ”SNR-Normalisation for Robust Speech Recognition”, Proceedings of ICASSP, Atlanta, Georgia, pp. 331-334, 1996.

[7] Claes, T., Xie, F., van Compernolle, D. ”Spectral Estimation and Normalization for Robust Speech Recognition”, Proceedings ICSLP, pages 1997-2000, 1996.

[8] 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.

[9] 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.

[10] Zeljkovic, I., Narayanan, S., ”Improved HMM Phone and Triphone Models for RealTime ASR Telephony Applications”, Proceedings ICSLP, pp. 1105-1108, 1996.

[11] Lee, C.-H., Juang, B.-H., Chou, W., Molina-Perez, J. J., ”A Study on Task-Independent Subword Selection and Modeling for Speech Recognition”, Proceedings ICSLP, pp. 1820-1823, 1996.