Robust Spelling and Digit Recognition in the Car: Switching Models and Their Like

Björn Schuller[1]*, Martin Wöllmer[1], Tobias Moosmayr[2], Gerhard Rigoll[1]

[1] *Technische Universität München, Institute for Human-Machine Communication, 80290 München, Germany*

[2] *BMW Group, Forschungs- und Innovationszentrum, Akustik, Komfort und Werterhaltung, 80788 München, Germany*

Introduction

Performance of speech recognition systems strongly degrades in the presence of background noise, like the driving noise in the interior of a car. We aim to improve noise robustness focusing on all major levels of speech recognition: feature extraction, feature enhancement, and speech modeling. Different auditory modeling concepts, speech enhancement techniques, training strategies, and model architectures are implemented in an in-car digit and spelling recognition task, which considers noises produced by various car types and driving conditions. Matched conditions training and auditory modeling techniques like Perceptual Linear Prediction (PLP) are applied in order to improve recognition rates. We prove that joint speech and noise modeling with a global Switching Linear Dynamic Model (SLDM) capturing the dynamics of speech, and a Linear Dynamic Model (LDM) for noise, outperforms speech enhancement techniques like Histogram Equalization (HEQ).

Speech Database

The digits "zero" to "nine" as well as the letters "A" to "Z" from the TI 46 Speaker Dependent Isolated Word Corpus [1] are used as speech database for the noisy digit and spelling recognition task. The database contains utterances from 16 different speakers - 8 female and 8 male speakers. For the sake of better comparability with the results presented in [2], only the words which are spoken by male speakers are used. For every speaker 26 utterances were recorded per word class whereas 10 samples are used for training and 16 for testing. Consequently the overall training corpus consists of 80 utterances per class while the test set contains 128 samples per class.

Noise Database

As interior noise masking varies depending on vehicle class and derivates, speech is superposed by noise of four different vehicles as listed in Table 1. Not only the vehicle type but also the road surface influences the characteristics of interior noise. Hence, three different surfaces in combination with typical velocities have been considered as shown in Table 2. The lowest excitation provides a driving over a smooth city road at 50 km/h and medium revolution (CTY). Thus at this profile noise caused by wind, engine, wheels etc. has its minimum. The subsequent higher excitation is measured at a highway drive at 120 km/h (HWY). In that case wind noise is a multiple higher than for a drive at 50 km/h. The worst and loudest sound in the interior of a car provokes a road with big cobbles (COB). At 30 km/h wind noise can be neglected but the rough cobble surface involves dominant wheel and suspension noise. Table 3 shows the mean SNR levels for all four car types at each driving condition.

Table 1: Considered vehicles

Vehicle	Derivative	Class
BMW 5 series	Touring	Executive car
BMW 6 series	Convertible	Executive car
BMW M5	Sedan	Exec. sports car
MINI Cooper	Convertible	Super-mini

Table 2: Considered road surfaces and velocities

Surface	Velocity	Abbreviation
Big cobbles	30 km/h	COB
Smooth city road	50 km/h	CTY
Highway	120 km/h	HWY

Table 3: Mean SNR levels for noisy speech utterances

Car Noise	SNR	Car noise	SNR
530i, CTY	-8 dB	645Ci, CTY	-3 dB
530i, HWY	-15 dB	645Ci, HWY	-13 dB
530i, COB	-23 dB	645Ci, COB	-19 dB
M5, CTY	-4 dB	Mini, CTY	-5 dB
M5, HWY	-11 dB	Mini, HWY	-15 dB
M5, COB	-21 dB	Mini, COB	-24 dB

In spite of SNR levels below 0 dB, the noisy test sequences are still well audible since the recorded noise samples are lowpass signals with most of their energy in the frequency band from 0 to 500 Hz. Consequently, there is no full overlap of the spectrum of speech and noise.

Apart from car noises (CAR), two further noise types are used in this work: first, a mixture of babble and street noise (BAB) at SNR levels 12 dB, 6 dB, and 0 dB, recorded in downtown Munich. This noise type is relevant for in-car speech recognition performance when driving within an urban area with open windows. Furthermore, additive white Gaussian noise (AWGN) has been used (SNR levels 20 dB, 10 dB, and 0 dB).

In order to simulate the worst case scenario which combines all three noise types, the speech utterances were also superposed with the car noise, babble and street noise, and AWGN at the same time (ALL) resulting in an overall SNR of -15 dB.

*Email:schuller@tum.de

Experiments and Results

For every digit an HMM was trained, whereas each HMM consists of 8 states with a mixture of three Gaussians per state. 13 Mel-frequency cepstral coefficients (MFCC) as well as their first and second order derivatives were extracted. In addition the usage of PLP features instead of MFCC was evaluated. Attempting to remove the effects of noise, various speech enhancement strategies were applied: Cepstral Mean Subtraction (CMS), Mean and Variance Normalization (MVN), Histogram Equalization, Unsupervised Spectral Subtraction (USS), and Advanced Front-End Wiener Filtering (AFE) [3]. However, as can be seen in Table 4, for stationary lowpass noise like the "CAR" and "BAB" noise types, the best average recognition rate can be achieved when enhancing the speech features using a global Switching Linear Dynamic Model [4] for speech and a Linear Dynamic Model for noise. For speech disturbed by white noise, the best recognition rate (93.3%, averaged over the different SNR conditions) is reached by the autoregressive Switching Linear Dynamical Model (AR-SLDS) introduced in [2], where the noisy speech signal is modeled in the time domain as an autoregressive process. This concept is however not suited for lowpass noise at negative SNR levels: for the "CAR" noise type a poor recognition rate of 47.2%, averaged over all car types and driving conditions, was obtained for AR-SLDS modeling.
In case an HMM recognizer without feature enhancement is applied, PLP features perform slightly better than MFCC.

Table 4: Mean isolated digit recognition rates for different noise types, noise compensation strategies, and features (training on clean data)

$\mathbf{Strategy_{feat.}}$	CAR	BAB	AWGN	ALL
SLDM_{MFCC}	99.5%	99.3%	87.8%	83.0%
HEQ_{MFCC}	98.2%	96.5%	77.5%	69.2%
CMS_{PLP}	97.7%	97.9%	72.7%	54.0%
MVN_{MFCC}	94.9%	93.3%	79.1%	54.5%
CMS_{MFCC}	97.0%	97.2%	72.2%	37.3%
HEQ_{PLP}	97.2%	95.3%	66.5%	34.5%
USS_{MFCC}	93.5%	92.3%	53.2%	40.3%
AFE_{MFCC}	87.9%	92.8%	64.1%	47.4%
none_{PLP}	81.1%	90.6%	67.7%	37.3%
none_{MFCC}	75.1%	88.4%	63.3%	32.8%
$\mathrm{AR{-}SLDS}_{none}$	47.2%	78.5%	93.3%	43.8%

Table 5 summarizes the mean recognition rates of an HMM recognizer without feature enhancement for three different training strategies: training on clean data, Mismatched Conditions Training, and Matched Conditions Training. Mismatched Conditions Training denotes the case when training and testing is done using speech sequences disturbed by the same noise type but at unequal noise conditions (SNR levels and driving conditions respectively). Matched Conditions Training means training and testing with exactly identical noise types and noise conditions.The best MFCC feature enhancement methods were also applied in the spelling recognition task (see Table 6). Again, for noisy test data, SLDM perform better than conventional techniques like HEQ.

Table 5: Mean isolated digit recognition rates of an HMM recognizer without feature enhancement for different noise types and training strategies: Matched Conditions (MC), Mismatched Conditions (MMC) and with clean data

Training	clean	CAR	BAB	AWGN
clean data	99.9%	75.1%	88.4%	63.3%
MMC	79.4%	96.9%	98.7%	68.5%
MC	99.9%	99.7%	99.7%	99.2%

Table 6: Mean spelling recognition rates for different noise types and noise compensation strategies (training on clean data)

$\mathbf{Strategy_{feat.}}$	clean	CAR	BAB	AWGN
SLDM_{MFCC}	92.7%	83.0%	81.6%	64.2%
HEQ_{MFCC}	91.8%	70.2%	69.4%	48.2%
CMS_{MFCC}	93.0%	73.8%	69.8%	47.1%
none_{MFCC}	91.0%	58.8%	66.6%	44.3%

Conclusion

The digit and spelling recognition task in this work examines various auditory modeling, feature enhancement, speech modeling, and training strategies for a wide range of different noise types. Thereby speech enhancement with an SLDM prevails for lowpass car noises, whereas AR-SLDS was proven to be the best technique for white noise. Mismatched Conditions Training is able to improve noisy speech recognition rates with respect to clean training. An upper border for the recognition performance is determined when using Matched Conditions Training, which assumes perfect knowledge of the noise properties.

Acknowledgment

We would like to thank Jasha Droppo and Bertrand Mesot for providing SLDM and AR-SLDS binaries.

References

[1] Doddington, G. R., Schalk, T. B.: Speech recognition: turning theory to practice. IEEE Spectrum (1981) 26–32

[2] Mesot, B., Barber, D.: Switching linear dynamical systems for noise robust speech recognition. IEEE Transactions on Audio, Speech and Language Processing (2007)

[3] Speech processing, transmission and quality aspects (STQ); Distributed speech recognition; Advanced front-end feature extraction algorithm; Compression algorithms. ETSI standard doc. ES 202 050 V1.1.5 (2007)

[4] Droppo, J., Acero, A.: Noise robust speech recognition with a switching linear dynamic model. Proceedings of the International Conference on Acoustics, Speech and Signal Processing (2004)