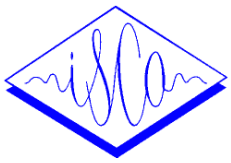

First International Conference on
Spoken Language Processing (ICSLP 90)
Kobe, Japan
November 18-22, 1990

PROBLEMS OF SPEECH RECOGNITION IN MOBILE ENVIRONMENTS

A. NOLL

aspect

Gesellschaft für Mensch-Maschine-Kommunikation mbh
Gutenbergring 38
D-2000 Norderstedt / Hamburg
Germany

ABSTRACT

For speech recognition in mobile environments a number of severe problems have to be solved. This paper gives an overview of several approaches dealing with the basic problem of noise handling in speech recognition in relation to the characteristica of noise in the mobile environment. Using a qualitative description of the noise characteristica, it can be seen that most of the currently investigated noise-handling strategies are just capable of handling parts of the phenomena of "mobile noise". From the analysis of these algorithms a general architecture of a noise resistent speech recognition strategy for further investigation is derived.

1. INTRODUCTION

In the area of speech recognition at least small- to moderate-size vocabulary systems have reached a level of recognition performance that is acceptable in many application areas. But some of the most promising applications require a level of noise robustness which has not yet been reached sufficiently.
One of these application areas is mobile radio, since it is expected that car telephones will become a significant factor of safety risks on our roads.
The reason for this is quite clear. Concentrating on the phone number, dialling by pushing the appropriate keys and checking this action on the telephone display distracts the drivers attention from the road. A speech-recognition and synthesis device enables the user at least to keep visual contact with the road. Additionally using features like 'hands-free' operation and dialling by names reduces the safety risks of mobile telephony nearly to the level of normal 'small talk' in the car. This motivation besides many others has led to an increased interest in robust speech recognition for such difficult application areas.

This paper tries to give an overview of currently investigated algorithms and strategies for speech recognition in noisy environments under the 'mobile environment' point of view. For this purpose a short qualitative characterization of the 'mobile noise' phenomena is given in Chapter 2. Chapter 3 contains a classification of several approaches dealing with the noise problem using mainly reports from ICASSP '90 to get a quite new state-of-the-art overview. Finally, from this analysis a promising 'noise resistent' speech recognition strategy for the mobile environment is derived.

2. NOISE IN MOBILE ENVIRONMENTS

In the following a qualitative characterization of those acoustic signals is given which have to be treated as noise for the recognition process. To be clear we define all signals in the environment to be noise, except those speech utterances which consist of the words of the recognition vocabulary.

For simplification we will focus on signals in a driving car, because their characteristics could easily be generalized for most other mobile environments.
The various noise signals in a car can simply be grouped using their production places. We differentiate between noise coming from inside and outside the car, noise produced by the car itself and noise produced by the movement of the car.

The first type of noise which comes from inside the car is produced either by the passengers, by audio equipment or both. It consists mainly of speech and music and includes features like large spectral variance over time and unpredictable signal-to-noise ratios (SNR) but depends just on the discipline of the user.
The second type includes all types of traffic noise and differs to a large extent in dependence on time and on the location of the car.

The best reproducible noise in terms of spectral behaviour and energy is produced by the car itself and by the movement of the car. This is caused by the engine and by air noise in dependence on speed, acceleration and the brand of the car but with slow changes in time with respect to the frame rate of a speech recognizer.

This simple classification of different acoustic signals which can be recorded in a car shows, that a speech recognition system has to handle an unpredictable mixture of different noise types with a possible large range of SNRs. Although the last type of noise has on the average the largest impact on the overal SNR, some correlations of the different types like high traffic noise at low engine noise in a city and vice versa on a highway forces a recognizer to be resistent against all of these noise types.
Finally, in dependence on the overall SNR of the noise mixture, the Lombard effect has to be taken into account.

3. NOISE HANDLING ALGORITHMS

In the following, several methods for decomposing a signal into a noise and a speech portion will be described. Investigations on most of them have been published in the latest literature (mainly ICASSP90).
We will focus on algorithms which do not dependent on special equipment e.g. microphone arrays. Since reports on experiments using different data are never comparable, no classification in terms of the quality of a certain implementation will be given.
It has also to be noted that most of the algorithms analysed in the following are not specially designed for the mobile environment as described in Chapter 2.

The first group of approaches deals with certain types of feature selection for noise robustness. In these approaches the noise is not modeled explicitly. The basic idea behind them is to emphasize certain spectral and temporal features of speech during signal analysis to increase the inherent SNR. This could be performed by trying to simulate the human auditory system [1] or using a priori knowledge of the spectral and temporal behaviour of speech explicitly [2][3].
The advantage of this group of approaches is, that no explicit assumptions are made for the noise signal. Therefore a large variety of different noise types can be handled. On the other hand no improvement could be expected if the noise consists of speech-like sounds (except the well-known improvement of recognition performance by using temporal features as such [4]).

In the second group of approaches no assumptions about the speech signal is made, but a priori knowledge of the noise is used to perform a decomposition. The main assumptions are that the noise is additive to the speech and can be described by a temporal stationary stochastic process. An often used and also simple algorithm in this group is noise subtraction using a noise pattern which can be adapted during the run-time of the system [5].
A in principle similar but concerning the initial estimation and the decomposition a more sophisticated method is described in [6]. This algorithm uses a Gaussian autoregressive model for the noise process of which the parameters are trained once. Another method which is similar from the basic assumptions but which estimates the noise distortions during the recognition process itself is given in [7], where the decomposition is integrated into the distance measurement. To some extent this is similar to [10], where decomposition is performed in the vector quantization (VQ) step of a hidden Markov model (HMM) with discrete emission probability density functions. Here the decomposition of speech and noise is performed by mapping the observed noisy vectors to clean speech VQ regions using trained probabilistic functions.
In a mobile environment this group of algorithms can improve recognition performance significantly for the noise signals which are slowly changing in time (e.g. noise from the car engine and noise produced by car movement) as has been shown in [5], if the noise model can be adapted sufficiently during recognition. They will gain less or no improvement for the other types of noise. Additionally, if the adaptation process is controled via signal energy, problems occur for low SNRs.

The last group of approaches integrates the modelling of noise and speech into the training procedure of the recognition systems. A good example of this group is a neural net system which is trained with noisy speech as input to produce clean speech as output like in [8]. In this paper special transformations for the hidden layer to the output layer are used to decompose the speech from the input signal.
A similiar idea but in a completely different system architecture is described in [10]. Here complete HMMs are trained for the noise signals under consideration to model the spectral and temporal structures. During recognition the search space is expanded by the noise HMM,

increasing it by a factor which is equal to the number of states of the noise models.
Training the noise signals seems to be the most general approach and therefore well suited to the noise problems of the mobile environment since this approach is at least theoretically capable to handle a large variety of different noise types.

4. CONCLUSIONS

In most of the mentioned papers no real data have been used for performance experiments and just one gives speech recognition results in a mobile environment. For simplification we will assume that similar restrictions and assumptions in certain algorithms would produce similar results on the same data in dependence on the complexity of the underlying models, so that we can just look at the different groups of algorithms in the following.

We denote the three basic types of noise as given in Chapter 2 with N1 for speech like sounds, N2 for unpredictable but mostly short distortions and N3 for permanent quasy-stationary noise. Lombard effects in very low SNRs will be denoted as L.

For the groups of algorithms we denote with A1 special signal-analysis approaches for speech modelling, with A2 noise modelling approaches, and with A3 general noise and speech training algorithms.
The following matrix gives an overview of the theoretical noise handling capability using a '+' sign for good, a 'o' for moderate and '-' sign for bad coverage for each combination of algorithm and noise. Of course the depicted weighting of capabilities can only be seen relatively using an optimal modelling of the underlying assumptions of each group and using the same recognition criterion:

	N1	N2	N3	L
A1	-	o	+	o
A2	-	-	+	-
A3	+	+	+	+

Table 1: Theoretical coverage of noise handling capability for the three groups of algorithms.

From this theoretical point of view we only can conclude that a complete integrated training of speech and noise in a recognition system has to be used. But from a practical point of view, it is nearly impossible to train all effects of the three noise groups in all possible variants and dependencies. If we take the trainability and additionally the complexity of implementation given the state-of-the-art signal-processing technology into account for implementing a real-time system in a mobile environment, the content of Table 1 changes as given in Table 2:

	N1	N2	N3	L
A1	-	o	o	o
A2	-	-	+	-
A3	-	-	o	o

Table 2: Practical coverage of noise handling capability for the three groups of algorithms.

Group A1 degrades for noise N3 since for today's algorithms it is much easier to model the noise of N3 in a large scale of SNRs than to model speech in N3 with acceptable performance.

Group A3 degrades drastically either because of the unreasonable high effort for training all effects in advance or because of the lacking of fast adaptation algorithms to prevent such a training.

The conclusion from this is quite obvious. A noise resistent speech recognition system for the mobile environment using state-of-the-art technology should be based on speech enhancing features and fast adaptable noise models to reduce the effort of an integrated training approach.

The recognition process should be based on a connected-word recognition algorithm since explicit endpoint detection in a noisy environment causes a large number of additional recognition errors. If accepted by the application, an isolated word syntax could be used to prevent most of the insertion and deletion errors. Furthermore in this case simple rejection techniques could be used.

A simplified information flowchart including the optimization criteria for the differnt models of such a system architecture is given in Fig. 1:

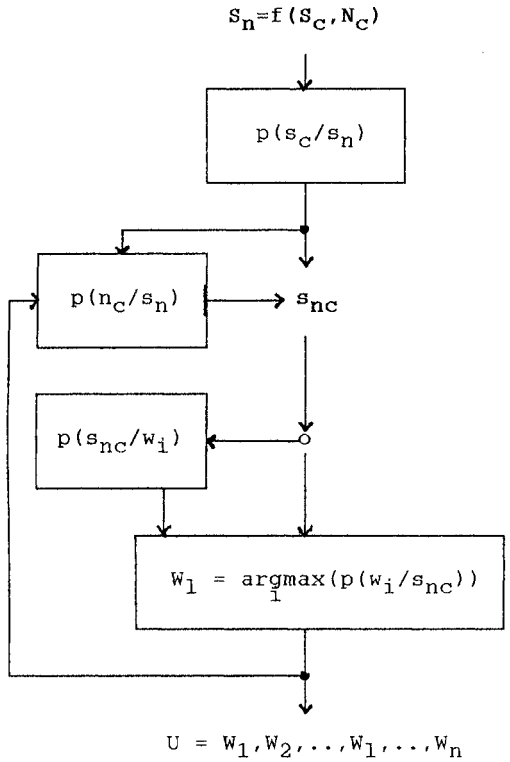

Fig. 1 Flowchart of system architecture

The incoming signal S_n is preprocessed using a clean speech and a pure noise model resulting in s_{nc}, which will be used for creating the word models during training as well as for recognizing the spoken word sequence.

Since the noise model focuses on the N3 type of noise, adaptation is performed using the segmentation of the recognition process to avoid adaptation problems. Whether the signal s_{nc} can be calculated using the speech and the noise model simultaneously has to be investigated. A first implementation of such a system architecture is described in [11].

REFERENCES

[1] TH. Sreenivas, K. Singh, R. Niederjohn, 'Spectral Resolution and Noise Robustness in Auditory Modelling', IEEE Proc. ICASSP'90, Albuquerque, pp. 817-820, April 1990

[2] T.F. Quatieri, R.J. McAuley, 'Noise Reduction Using a Soft-Decision Sine-Wave Vector Quantizer', IEEE Proc. ICASSP'90, Albuquerque, pp. 821-824, April 1990

[3] B.A. Hanson, T.H. Applebaum, 'Robust Speaker-Independent Word Recognition Using Static, Dynamic and Acceleration Features: Experiments with Lombard and Noisy Speech', IEEE Proc. ICASSP'90, Albuquerque, pp. 857-860, April 1990

[4] H.Ney, 'Experiments on Mixture-Density Phoneme-Modelling for the Speaker-Independent 1000-Word Speech Recognition Darpa Task',IEEE Proc. ICASSP'90, Albuquerque, pp. 713-716, April 1990

[5] A. Noll, H.H Hamer, H. Piotrowski, H.W. Rühl, S. Dobler, J. Weith, 'Real-Time Connected-Word Recognition in a Noisy Environment', IEEE Proc. ICASSP'89, Glasgow, pp. 679-682, May 1989

[6] Y. Ephraim, ' A Minimum Mean Square Error Approach for Speech Enhancement', IEEE Proc. ICASSP'90, Albuquerque, pp. 829-832, April 1990

[7] A. Ereil, M Weintraub, ' Estimation Using Log-Spectral-Distance Criterion for Noise-Robust Recognition', IEEE Proc. ICASSP'90, Albuquerque, pp. 853-856, April 1990

[8] S. Tamura, M. Nakamura, 'Improvements to the Noise Reduction Neural Network', IEEE Proc. ICASSP'90, Albuquerque, pp. 825-828, April 1990

[9] H. Gish, Y. Chow, J.R. Rohlicek, 'Probabilistic Vector Mapping of Noisy Speech Parameters for HMM Word Spotting', IEEE Proc. ICASSP'90, Albuquerque, pp. 117-120, April 1990

[10] A.P. Varga, R.K. Moore, 'Hidden Markov Model Decomposition of Speech and Noise', IEEE Proc. ICASSP'90, Albuquerque, pp. 845-848, April 1990

[11] H.W. Rühl, S. Dobler, P. Meyer, A. Noll, H.H. Hamer, H. Piotrowski, 'Speech Recognition in the Noisy Car Environment', to be published in Speech Communication