OPTIMIZATION OF A SPEECH RECOGNIZER FOR AIRCRAFT ENVIRONMENTS

Volker Schless *Fritz Class* *Peter Sandl**

DaimlerChrysler AG, Research and Technology, Wilhelm-Runge-Str. 11,
D-89081 Ulm, Germany
e-mail: {volker.schless,fritz.class}@daimlerchrysler.com
*DaimlerChrysler Aerospace AG

10.21437/Eurospeech.1999-632

ABSTRACT

Speech recognition in aircrafts can greatly simplify operation of equipment in both military and civil environments. This paper describes the development of specialized recognizers for two military applications: One for assisting a jet pilot wearing a breathing mask and another for a radar evaluator within an aircraft (similarly to AWACS). In both cases it is not practical to collect sufficient speech data under real conditions for a robust specialized recognizer. This paper describes two methods for overcoming this data problem and building a recognizer with minimal effort: retraining the baseline system with real application data and combining the baseline system with a new trained system. These methods greatly improved the performance of the baseline system (US English recognizer adapted to car environment).

1. INTRODUCTION

Speech recognition is required in various environments. Depending on the application, conditions may differ significantly with respect to noise, speaker, and input equipment. To achieve high performance in various environments the models have to be trained accordingly. Therefore, a large number of training samples is often necessary. However collecting speech data in real environments is time consuming. Thus it is not possible to completely rebuild recognizers for each application. Especially for aircraft environments, costs are extremely high and real speech samples are very rare. Sometimes even the simulation of environments is very cost intensive. For example, the commercial DaimlerChrysler recognizer for the car environment is based on about 400,000 sentences of speech. So renewed data collection for each new environment is not feasible.

Nevertheless our goal is to achieve recognition rates comparable to the case in which training and testing conditions are matched. This is assumed to be the upper bound of performance. One possibility to approximately obtain this goal is to apply robust methods for speech processing to an unadapted system. In the literature speech recognition in aircraft conditions is performed with many different techniques. Tests for the jet environment (F 16) include, for example, LDA for preprocessing [8], spectral subtraction [9], cepstral subtraction [10] and parallel model combination (PMC) [4, 5]. [3] proposes the joint application of PMC and spectral subtraction.

Contrary to these methods we try to adapt a speech recognizer which is robust under one set of conditions to a new environment with incomplete training data. In this paper the existing recognizer (baseline system) was originally designed and trained for the car environment. The goal is to create a system which is robust under aircraft conditions. Only a small amount of data for two different applications is available for adaptation. The two target environments are a military pilot in a cockpit speaking through a breathing mask and a radar evaluator sitting in an aircraft with ambient noise.

This paper describes two methods for constructing a speech recognizer in adverse environments. First, training a new recognizer with a small amount of data and combining its characteristics with the characteristics of the baseline system. Second, retraining the baseline system. Both methods were tested for two different environments and compared to the performance of the baseline recognizer.

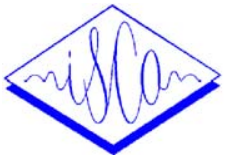
2. METHODS

First we describe two methods for simulating the adverse environments. Various strategies have been investigated for recoding depending on the kind of noise. Afterwards the training methods are described.

2.1. Simulating Noisy Environments

Basically there are two methods for simulating adverse environments. First, adding noise to existing clean speech samples. Of course this is possible only if the noise is actually additive in the time domain and effects of speaker adaptation to the noise like the Lombard reflex [6] are negligible. Second, re-recording speech data under the conditions in question. We only use the second method for our work here because recordings are more realistic. This increases the effort compared to simply adding noise to the existing speech samples because recordings have to be made separately for each environment.

In this work we follow two different methods for obtaining realistic data. In the case of the radar environment, background noise can be assumed stationary. An example of



6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

a typical speech recording is illustrated in Figure 1. Here the simulated conditions were produced by introducing the noise through the loudspeakers. The noise level is 85 db.

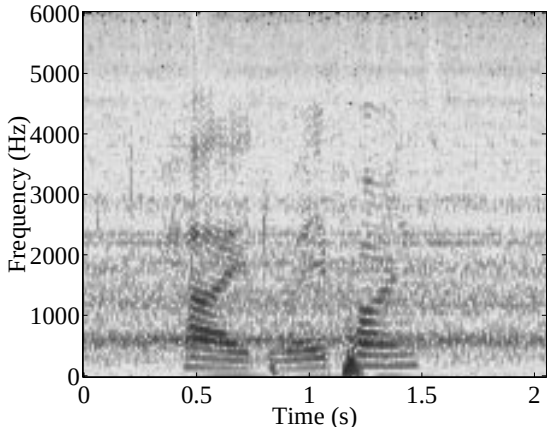


Figure 1: Spectrogram of speech with background noise for the radar environment (utterance "hide waypoint")

For the jet environment the situation is completely different. Here the pilot uses a breathing mask. The noises from outside are well damped because the microphone is located inside the mask. So simulating jet noise was considered irrelevant. However, breathing noises occur quite often. The characteristics of this noise can be seen at the end of the utterance in Figure 2. To simulate these conditions, speakers had to wear a breathing mask while producing the required utterances for training and testing.

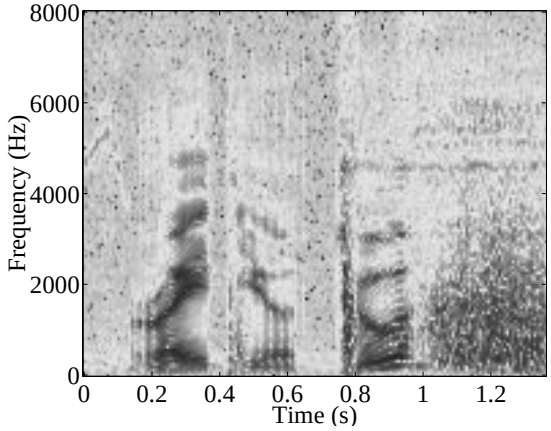


Figure 2: Spectrogram of speech with breathing noise (utterance "radio 2")

2.2. Training with Noisy Speech

Depending on the noise characteristics two different ways of incorporating noise into the training process will be used. In the case of stationary noise nothing special needs to be said. Since the noise corrupts all utterances equally, training can be carried out in the same manner as for clean conditions. In the case of intermittent breathing noise, breathing can be explicitly transcribed. Specialized breathing models were introduced for breathing in general and for breathing in and breathing out. Because transcribing the breathing noise of each utterance is very time consuming, tests were made with explicitly transcribed breathing noise and with normal transcription where breathing is assumed to occur at the beginning and the end of each utterance. Experiments show whether the effort of transcribing breathing noise has a positive effect on system performance.

2.3. Retraining of the Baseline System

Two different methods for adapting the recognizer to the new environment are applied. First, the system can be retrained with the limited amount of training data available from the new environment. Here additional training iterations on the HMM models of the baseline recognizer are necessary. Since HMM models of the baseline recognizer are used for initialization, training results will be different from the results when starting with equal distributed model statistics. Models which do not occur in the training set remain unchanged. Results can be obtained very fast with this procedure because only the final training step has to be repeated and the amount of training data is very small.

2.4. Combining HMM Probabilities

HMM characteristics of the baseline system can also be combined with models of a system that is based on the few speech samples of the new environment. To do this, it is necessary to build a new recognizer with the new speech data. The training procedure and the details of the new and the baseline systems have to be identical in order to be able to combine them. Once the HMM probabilities for the new models have been computed a weighted sum of these probabilities and the probabilities of the baseline models is computed. These weighted sums form the new combined models. The weighting factor is purely heuristic.

3. EXPERIMENTS AND RESULTS

In this section details of the training and testing environment are explained. Afterwards we present results for the two adverse environments and the methods, proposed above.

3.1. Training and Testing Issues

The baseline system (US English recognizer for the car environment) is based on about 400,000 training sentences. LDA-transformed mel-cepstrum features are used (see [2, 7] for details) along with spectral subtraction [1].

Training and testing is carried out separately for the radar application and the jet application. For the radar environment 3372 utterances with 42 different commands were recorded. The vocabulary contains 144 command words including letters and digits. For testing, 321 additional commands from 4 speakers were used. Although all recordings were made in English, most speakers were German. The average signal-to-noise ratio of the training and test set is 13 db.

The training set for the jet environment consists of 3960 utterances with 42 different commands. This vocabulary differs from the radar application. It includes 138 commands, digits, and letters. For testing, only 29 commands were available that were recorded under real jet conditions. Because this test set is not sufficient for extensive experiments, also tests on the training set were performed. To cover breathing noises caused by the oxygen mask, we introduced additional noise models. These describe either breathing in general or breathing in and breathing out separately.

The command recognition rate is the performance. A command is considered correctly recognized only if all of its components are correctly recognized.

3.2. Results for the Radar Application

The methods described here were tested first for the radar environment. The recognition rate using the unadapted car recognizer (baseline system) was only 68% (see line 1 in Table 1). After training the recognizer with the new data is was combined with the baseline recognizer in a ratio of 9 (baseline) to 1 (new). This increased the recognition rate to 77%. Increasing the weight factor to 2 (baseline) : 1 (new) yielded a recognition rate of 86%. Further increasing it to 1:1 yielded 89%. This result differs not significantly from the performance of the radar recognizer alone (92%). Though the influence of the models of the baseline system may lead to enhanced robustness in slightly different and changing environments.

The last line of Table 1 shows the result of the retrained recognizer which is obtained by adding training iterations with the new speech samples to the models of the car recognizer. Compared to the equal weighting method, retraining shows no difference.

Weighting ratio of HMM models car : radar	recognition rate
car recognizer (baseline)	68%
9 : 1	77%
2 : 1	86%
1 : 1	89%
1 : 2	90%
radar recognizer	92%
retrained car recognizer	89%

Table 1: Command recognition rates for the radar application with different recognizers

3.3. Results for the Jet Application

The recognition rate using the car recognizer in the jet environment was 37% (see line 1 of Table 2). The decrease in performance is much greater than for the radar environment. Retraining the car recognizer with speech data from the jet environment increased the recognition rate to 87.6%. A recognizer trained exclusively with the new data is not significantly better (lines 2 and 3 in Table 2). Now additional breathing models were used. In the first experiment breathing was assumed present at the beginning and the end of each utterance and was not explicitly transcribed. This increased the recognition rate from 88.3% to 89.5%. Explicit transcription of breathing did not improve any further. Tests with two different breathing models (breathing in and out) also yielded no improvement.

The last two lines in Table 2 show the effect of weighting compared to the pure models above. Equal weighting leads to no degradation in performance compared to the systems that were trained exclusively in the jet environment. We expect however, that weighting gives an enhanced robustness in slightly different conditions. Contrary to the radar environment, retraining the baseline recognizer with new speech samples leads to degraded performance compared to combining characteristics of the two recognizers (see lines 2 and 6 of Table 2).

Finally tests with 29 utterances recorded in real jet conditions were performed for validation of the previous results. Table 3 shows recognition rates for 4 different recognizers. Because the test set is very small a significant improvement can only be noticed between the car recognizer and the other 3 systems.

4. CONCLUSION

In this paper we presented experiments of two methods for adapting a speech recognizer to adverse environments of two real applications. Only a small amount of training data was available. Weighting statistics of a recognizer trained with this data and a baseline system shows better results than retraining the baseline system for the jet environment. For the radar application these two methods achieve comparable recognition rates. In both cases the underlying statistics of the baseline system constructed from many speech samples may lead to enhanced robustness of the adapted recognizer in real environments.

REFERENCES

1. S. Boll. Suppression of acoustic noise in speech using spectral subtraction. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 27(2):113–120, 1979.

2. F. Class, A. Kaltenmeier, and P. Regel-Brietzmann. Optimization of an HMM–based continuous speech recognizer. In *Proc. European Conf. on Speech Communication and Technology*, pages 803–806, Berlin, Germany, 1993.

recognizer	recognition rate
car recognizer (baseline)	37.0%
retrained car recognizer	87.6%
jet recognizer	88.3%
jet recognizer with default breathing model	89.5%
jet recognizer with exactly transcribed breathing model	89.1%
weighting car/jet recognizer 1:1	89.1%
weighting car/jet recognizer with default breathing model 1:1	89.5%

Table 2: Command recognition rates for the jet application with different training methods and breathing models

recognizer	recognition rate
car recognizer (baseline)	41%
jet recognizer with default breathing model	76%
jet recognizer with exactly transcribed breathing model	72%
weighting jet/car recognizer with default breathing model 1:1	79%

Table 3: Command recognition rates for the jet application with real data

3. J. Flores and S. Young. Adapting a HMM–based recogniser for noisy speech enhanced by spectral subtraction. In *Proc. European Conf. on Speech Communication and Technology*, volume 2, pages 829–832, 1993.
4. M. Gales and S. Young. HMM recognition in noise using parallel model combination. In *Proc. European Conf. on Speech Communication and Technology*, volume 2, pages 837–840, 1993.
5. J. Hung, J. Shen, and L. Lee. Improved parallel model combination techniques with split Gaussian mixtures for speech recognition under noisy conditions. In *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, volume 1, 1999.
6. J. Junqua. The influence of acoustics on speech production: A noise–induced stress phenomenon known as the Lombard reflex. *Speech Communication*, 20:13–22, 1996.
7. V. Schless and F. Class. SNR–dependent flooring and noise overestimation for application of spectral subtraction and model combination. In *Proc. Int. Conf. on Spoken Language Processing*, volume 4, pages 1495–1498, 1998.
8. O. Siohan, Y. Gong, and J. Haton. A comparison of three noisy speech recognition approaches. In *Proc. Int. Conf. on Spoken Language Processing*, pages 1031–1034, 1994.
9. H. van Hamme. ARDOSS: Autoregressive domain spectral subtraction for robust speech recognition in additive noise. In *Proc. Int. Conf. on Spoken Language Processing*, volume 2, pages 1019–1022, 1994.
10. C. Wu, V. Nguyen, H. Sabrin, W. Kushner, and J. Damoulakis. Fast self–adapting broadband noise removal in the cepstral domain. In *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, volume 2, pages 957–960, 1991.