KTH Speech, Music and Hearing

Speech recognition in the JAS 39 Gripen aircraft - adaptation to speech at different G-loads

Christine Englund

Supervisor:	Kjell Elenius
Approved:	11th March 2004
Examiner:	Kjell Elenius

....

(signature)

Centre for Speech Technology

Stockholm
11th March 2004

Master Thesis in Speech Technology

Department of Speech, Music and Hearing
Royal Institute of Technology
S-100 44 Stockholm

KTH Tal, musik och hörsel

Examensarbete i Talteknologi

Taligenkänning vid olika G-belastningar i JAS 39 Gripen

Christine Englund

Godkänt:
11th March 2004

Examinator:
Kjell Elenius

Handledare:
Kjell Elenius

Abstract

During the years 2000 and 2001, the speech of eight pilots flying the JAS 39 Gripen aircraft was recorded at Saab in Linköping. Thirty sentences were recorded on the ground and at different G-loads, with a maximum of 8G. The sentences were spoken in both Swedish and English. There was also some additional material which consisted of mostly spontaneous speech. The purpose of this Masters Thesis was to perform speech recognition on the material, and also adaptation to different G-loads, in order to see if this improved recognition. A reference test using a different clean speech database showed that the monophone model sets used for these experiments gave good performance in clean conditions.

A first recognition test on the whole Gripen material recieved word accuracies of below 15 % for both languages. These low results are likely to be due to many factors: the noisy conditions under which the material was recorded; the recording channel; the oxygen mask breath noises and the spontaneous speech. Furthermore, speech recognition deteriorates with changes in articulation and in voice characteristics, caused by background noise and high G-loads. A decreasing tendency in the recognition results could be observed with increasing G-loads. The ground condition recieved a word accuracy of 60 %, whereas speech in 4G recieved word accuracies of 40 % for Swedish and 30 % for English.

A re-training procedure was used to adapt the models to the characteristics of the speech. Furthermore, adaptation to specific G-loads was carried out using a combined MLLR and MAP approach. In the adaptation procedures, models for breath noise were trained from the adaptation material and included in the model sets in the subsequent tests. It was found that all recognition results improved significantly after adaptation, the ground condition now recieving scores of above 80 % for Swedish and just below 80 for English. For 4G, these figures were approximately 70 and 60 %. Tests on speech in the ground condition with and without the breathing models showed that modeling these acoustic events had a positive effect on the performance.

It can be concluded that high G-loads complicate speech recognition, and that adaptation significantly improved results. Furthermore, adding breath models resulted in improved recognition scores.

Sammanfattning

Under år 2000 och 2001 spelades en databas innehållande tal från åtta svenska piloter in på Saab i Linköping. Syftet var att erhålla ett material för forskning och utveckling inom området automatisk taligenkänning i flygplan. Piloterna satt vid inspelningstillfället inuti JAS 39 Gripen. Ett trettiotal fraser lästes, både på svenska och engelska. Inspelningar gjordes både på marken och i luften vid olika G-belastningar, maximalt 8G.

Syftet med det här examensarbetet var att använda databasen till taligenkänningsförsök, och att se vad som kan uppnås genom att adaptera igännkänningssytemet till olika G-belastningar.

Referenstester på tal inspelat i kontorsmiljö visade att de svenska och engelska monofonmodeller som användes i de här försöken gav goda resultat på ostört tal. Ett första igenkänningsförsök på hela Gripen-materialet gav däremot en ordnoggrannhet på mindre än 15 % för båda språken. Orsaker som har bidragit till de låga resultaten är inspelningskanalen (mikrofon placerad inuti andningsmasken), tryckandningen, bakgrunds-störningarna vid höga G-belastningar, samt förekomsten av spontantal i materialet. Också artikulation och röstkaraktär påverkas av buller och höga G-belastningar och försvårar taligenkänning.

Igenkänningsförsök på tal vid specifika G-belastningar visade att resultaten sjönk kraftigt med stigande G-belastningar. De bästa resultaten, cirka 60 %, erhölls för tal i tyst markmiljö, medan tal vid 4G:s belastning endast gav 40 respektive 30 % i ordnoggrannhet för svenska respektive engelska.

Två typer av adaptering utfördes. I ett omträningsförfarande adapterades taligenkänningsmodellerna till ett stort material som inkluderade tal vid ett flertal G-belastningar. Adaptering gjordes också till tal vid specifika G-belastningar. Här användes adapteringsmetoderna MLLR (Maximum Likelihood Linear Regression) och MAP (Maximum a posteriori). Akustiska modeller för tryckandning tränades under adapteringen och användes i påföljande tester.

Försöken visar att alla resultat förbättrandes markant efter adaptering, särskilt vid adaptering till specifika testförhållanden, där förbättringar på uppemot 100 % kunde observeras. Marktillståndet erhöll nu en ordnoggrannhet på cirka 80 % för båda språken, medan resultaten för tal i 4G steg till 70 % för svenska, och 60 % för engelska. Också införandet av andningsmodeller hade en påtagligt förbättrande effekt på resultaten. I det engelska fallet förbättrades resultatet från strax över 30 till strax under 60 %.

I praktiken krävs dock en bättre igenkänning. Bättre akustiska modeller, någon form av brusreducering, adaptering till den enskilde talaren, och framförallt en språkmodell, skulle bidra till bättre resultat.

Acknowledgments

I would like to thank my supervisor Kjell Elenius. Thanks also to Daniel Elenius and Giampiero Salvi for good suggestions and for taking the time to answer my questions.

Contents

1 Introduction

This thesis is part of the Master of Engineering program in Media Technology at the KTH (Royal Institute of Technology). The work was performed at the Department of Speech, Music and Hearing during the fall of 2003 and early 2004.

The topic of this thesis is speech recognition in military aircraft. During 2000 and 2001, a speech database was recorded at Saab in Linköping. The recordings were made in the JAS 39 Gripen aircraft with eight professional pilots. This thesis uses the database for speech recognition and adaptation experiments.

1.1 Problem formulation

For reasons that will be discussed later in this report, speech recognition on the Gripen material could be expected to be problematic. The purpose of this project was to perform speech recognition experiments on the material, and to see to what extent results could be improved by adapting the system to specific G-loads.

1.2 Tools

The experiments performed in this work used tools of the HTK toolkit (Hidden Markov Model Toolkit) (Young, 2002). HTK is designed for building HMMs (Hidden Markov Models) and optimized for building speech processing tools. Scripts to automize training and test procedures were written in the programming language Perl.

1.3 Method

For adaptation purposes, both a re-training procedure and a combined MLLR and MAP approach was used (Young 2002). Adaptation was performed on a large material as well as on smaller sets at specific G-loads.

The material contained non-speech acoustic events such as oxygen mask breath noises. These were shown to cause problems for the recognizer, and for this reason, new models for breathing were trained.

1.4 Outline of the report

The report starts out with a brief introduction to the theory of automatic speech recognition. This is followed by a chapter on speech recognition in aircraft. The advantages of such systems are explained, and research within the field is outlined. Chapter four describes the composition and characteristics of the data used for this project, and the next chapter discusses the initial recognition tests. In chapter six, the adaptation procedures are presented, and chapter seven deals with the reference tests that were also a part of this work. The obtained results are given in the results section, and are discussed in the subsequent chapter. Finally, the conclusions are stated in the last section of this report.

2 The theory of automatic speech recognition

Speech recognition is the technology that makes it possible for a computer to identify the components of human speech. The process can be said to begin with a spoken utterance being captured by a microphone, and to end with the recognized words being output by the system. The steps involved in this process will be described in the sections below.

Research in speech recognition began somewhere in the late 1940's and the rapid progress in computer technology has been of great importance to the development of this field. Today, an increasing number of products incorporate speech technology. Some of these are found in industry. In situations where the hands and eyes of persons are engaged elsewhere, voice control can be a great advantage. Other applications are booking systems over the telephone, in which case speech is a more intuitive means of communication than pressing buttons. Yet other applications include various forms of aids for the disabled, dictation systems, as well as toys and games.

Ideally, a speech recognizer would be able to accurately identify all possible words spoken by any person in any environment. In reality, however, system performance depends on a number of factors. Large vocabularies, multiple users, as well as continous speech (as opposed to words spoken in isolation) are factors that complicate the task of recognition. The same is true for speech in noisy environments.

2.1 Feature extraction

Today, signal processing is done in the digital domain, almost without exclusion. Prior to any such processing, the signal is sampled. In sampling, the signal is measured at certain equally spaced points in time. According to the sampling theorem, any bandwith-limited signal can be perfectly reconstructed if the sampling frequency, Fs, is at least twice the highest frequency represented in the signal (Alan and Willsky, 1997). The signal is also quantized with respect to its amplitude, and the quantization error, or noise, is determined by the number of bits used.

For ASR (Automatic Speech Recognition) applications, time domain representation of the signal is sub-optimal; a more compact and useful representation is desirable. Feature extraction, also called front-end analysis, is the process by which the acoustic signal is converted into a sequence of feature vectors. Several feature sets can be used for the vectors, MFCCs (Mel Frequency Cepstral Coefficients) are among the more common. It is desirable that the features (Martens et al., 2000):

- Allow an automatic system to discriminate between speech sounds that are similar sounding
- Allow for models to be created without an exessive amout of training data
- Supress characteristics of the speaker and the environment

In parametrizing the waveform, the time domain signal is treated as a sequence of frames, each of which is represented by a feature vector. The duration of the frames are commonly 25 ms, which is short enough to be assumed to come from a stationary process. Each frame is first multiplied by a window function, often a Hamming or Hanning window. This is done to smooth the edges of the frames, which would otherwise cause high frequency components to appear in the spectrum. The frequency content of the frames is then analysed, usually with FFT-analyses (Fast Fourier Transform). MFCCs are computed by taking the inverse Fourier transform of the logarithm of the amplitude spectra.

2.2 Classification

The transformation of speech into feature vectors is followed by the process of recognizing what was actually spoken. There are several approaches to this problem. A brief description of the main ones will be given here. These include: knowledge-based approaches, template matching, stochastic approaches and connectionist approaches. These methods are not mutually exclusive.

2.2.1 Pattern matching techniques

A pattern matching system is based on the idea of comparing input utterances to a number of prestored templates, i.e example acoustic patterns. Usually each template corresponds to a word in the vocabulary. The classifier will calculate the acoustic difference between the input utterance and each of the stored templates and choose the template which shows the highest acoustic similarity to the input.

Dynamic programming, a mathematical technique which finds the best non-linear match between the timescales of the two utterances, is often used in conjunction with pattern matching to compensate for variations in speed of the two spoken utterances.

Pattern matching techniques were widely used in commercial products in the 1970s and 1980s, but have since then become more or less replaced by more powerful methods (Holmes, 2001).

2.2.2 Neural networks

Neural networks are an attempt to model certain properties of the human nervous system. A network consists of a large number of nodes. These nodes are organized in layers and inter-connected with weights of different strengths. Information is fed to an input layer, processed by the net, and then fed to a layer of output units. The response of each node is generally determined by a non-linear function of the weighted sum of its inputs.

The network's abililty to correctly classify the input depends on the values of the weights and the optimal values are found during training. In training, some acoustic information, e.g. spectral amplitudes, is supplied to the input nodes of the network, and the output value is compared to the desired value, e.g. a phoneme. The error, the difference between the desired and the actual output, is used to modify the weights of the network. This process is repeated several times for each training utterance, to increase the likelihood of correct classification.

Although an interesting and promising technique, neural networks have not yet achieved success as a complete system for the recognition of continuous speech. Their strength lies mainly in classification, and they may be used for this purpose, serving as a component of larger HMM based systems (Holmes, 2001).

2.2.3 Knowledge-based approaches

Knowledge-based systems use knowledge in some form to distinguish between different types of speech sounds. In the late 70's and 80's, interest was taken in the so called expert systems, which base classification on rules formed from knowledge about the speech signal. To capture the great variability of speech, a large set or rules is required (Blomberg and Elenius, 2003).

Other types of systems are those oriented towards the human speech production process. Here the rules are instead specified in articulatory terms. In this case, classification is made by comparing synthesized speech with an unknown utterance. Although a potentially flexible and interesting technique, insufficient knowledge, and other difficulites, limit the usefulness of such systems as of yet.

2.2.4 Hidden Markov models (HMM)

Hidden Markov models are a powerful statistical method for modeling speech signals, and they are the dominating approach in speech recognition today.

A Hidden Markov model represents a language unit, for instance a word or a phoneme. It has a finite number of states and the transitions between these are probabilistic and take place once every time unit (a model may also remain in the same state). Each state has a probabilistic output function which represents a random variable or a stochastic process (Rabiner and Juang, 1986). Gaussian distributions are a common choise for representing these functions, and in reality, mixtures of Gaussians with individual means and variances and mixture weights are often used, as these allow any arbitrary function to be approximated.

When presented with an observation sequence, a model can determine the probability of having generated the observations, but since the observations do not uniquely define a particular state sequence, it is not possible to know which states were active, and in what order. This is why the process is said to be hidden.

The transition probabilities and the probability distributions along with their weights, are the parameters of an HMM. During training, these are optimized with respect to the training data, to increase the likelihood of the models having generated the data.

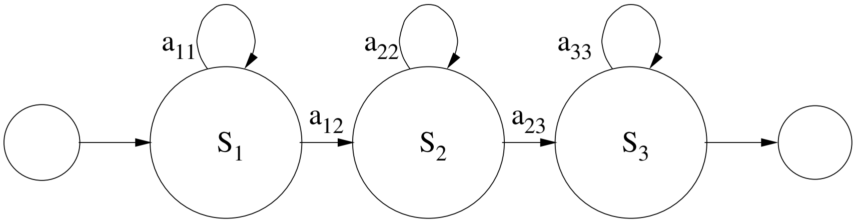

Figure 1: A simple three state Markov model with transition probabilities a_{ij}

2.3 Adaptation

The performance of automatic speech recognition systems can drop considerably when there is an acoustic mismatch between training and test data. The mismatch can be due to factors such as environmental noise, inter-speaker variability, and the acquisition channel.

Instead of training new models for the new condition, which would involve the costly process of collecting and preparing new speech data, adaptation techniques are often used. Adaptation is a way of using just a small amount of data to tailor existing models to the characteristics of, for example, a new speaker or a new environment.

Adaptation techniques can be divided into different modes. In *supervised* adaptation, the associated transcriptions are known, whereas *unsupervised* adaptation refers to situations where the adaptation data is unlabelled.

In *static* adaptation, the data is available in one block. *Incremental* adaptation, on the other hand, is done incrementally, as more data becomes available (Young, 2002).

2.3.1 Maximum likelihood linear regression (MLLR)

Maximum likelihood linear regression is an adaptation technique which applies linear transformations to clusters of acoustic units. The transformations are estimated from the adaptation data and are used to alter the means and variances of the Gaussian mixtures, so that these have a higher likelihood of having generated the observations.

MLLR is an example of an indirect adaptation technique; since data is clustered, all units are updated, even if they lack representation in the adaptation data. This makes MLLR effective for small amounts of data, but also leads to a quick saturation in performance when the amount of data increases (Siohan et al., 2001).

2.3.2 Maximum a posteriori (MAP)

Maximum a posteriori is another adaptation technique, which combines prior knowledge about the model parameters with information obtained from the adaptation data.

Contrary to MLLR, MAP is a direct adaptation technique in that it updates the acoustic units individually. Acoustic components not present in the adaptation data will not be updated. This means that MAP is not an ideal technique for small amounts of data. On the other hand, due to the detailed update of every component, it outperforms MLLR when more data is available.

The two described methods can be combined to improve results even further. In this case, the MLLR estimates can be used as prior information for MAP (Young, 2002).

3 ASR in aircraft

3.1 Advantages of ASR in aircraft

Cockpits are becoming more and more complex (Gordon, 1990), and they contain a multitude of control devices. This complex and often stressful environment poses high demands on the pilot's attention. Implementing speech technology as an alternative control strategy in aircraft would free the hands and eyes of the pilot, thus reducing the workload and better allow him or her to concentrate on the task.

Furthermore, functionality can be added to an ASR system on the software level, which is an advantage compared to installing new buttons and wires into the cockpit, where space is limited. This could also be beneficial from an economical point of view (Hammar, 1995).

3.2 Adverse conditions in the cockpit

Whereas a human being has a good ability to distinguish speech from noise, the performance of a speech recognizer degrades rapidly in adverse conditions, and especially when there is an acoustical mismatch between the training and test data.

The environment in the cockpit of a military aircraft is a challenging one for speech recognizers, due to environmental noise and changes of the pilot's articulation.

3.2.1 Sources of noise in the cockpit

The most common sources of noise are the aircraft engines, wind, environmental control systems, oxygen mask breath noise and electrical channel noise (Williamson, 1997). The noise level of a moderate fighter cockpit may well exceed 100 dBA. Methods of tackling noise include training in the noisy environment and noise cancellation.

Oxygen mask breath noise is a unique problem in aircraft speech recognition applications. A paper by Williamson (1997) describes two ways of dealing with this type of noise. One is to train the system with the oxygen mask, in which case the breath noise will be incorporated into the word models, minimizing its impact. Another, more flexible approach that has been shown to be successfull is the creation of separate models for breathing.

3.2.2 Speech variability

Speech variability is caused by G-loads, stress, fatigue and Lombard speech. G-loads occur when the aircraft accelerates or changes direction, and are expressed as units of the normal acceleration of gravity on earth, 9.8 km/s^2. In the pilot, recurrent and sustained high G-loads lead to fatigue and performance degradation (Vasiletz and Yakimenko, 1995). Affected bodily functions are for instance the blood flow, which also causes changes in the pumping rate of the heart and the oxygen supply to the brain. When subjected to extreme G-loads, the pilot may experience visual disturbances (greyout and blackout), and eventually loss of consciousness (occurs at 5-6G unless the pilot is protected).

The *Lombard effect* is the tendency to increase the vocal effort to compensate for loud background noise (Huang et al., 2001). When comparing normal and Lombard speech, changes in voice characteristics such as pitch, center of gravity, spectral tilt and formant frequencies can be observed (Stanton et al., 1988). Also the style of articulation and speaking rate may be modified by the speaker in these conditions.

3.3 Studies on ASR in aircraft

Much of the research on speech recognition in aerospace concerns military applications, wherefore results are often classified and hard to come by.

However, a paper by Williamson et al. from 1996 describes two flight tests carried out by the Pilot-Vehicle Interface Branch (FIGP) of Wright Laboratory (WL) (Williamson et al., 1996). The objective of the first experiment was to measure word recognition accuracy of the ITT VRS-1290 speech recognition system on the ground and in 1G and 3G flight conditions. The ITT VRS-1290 system is speaker dependent and handles continuos speech. The experiment used a NASA Lewis Research Center OV-10A aircraft and the speech of 13 professional pilots.

Subjects were tested in five conditions; in the laboratory, on the ground with engines off, in 1G, in 3 G and once more in 1G to test possible fatigue effects. The laboratory tests involved template training for the voice of each speaker. Aircraft test sessions involved noise calibration, i.e. generating templates for the background noise, later used to adjust the voice templates for use in higher noise conditions. The vocabulary consisted of 54 words, and the tests used a syntax describing permitted word sequences.

The mean word accuracies for the 13 speakers were 98.24 % in the laboratory, 98.42 % in the hangar, 98.55 % in 1G, 97.3 % in 3G and 98.15 % for the repeated test in 1G. A performance degradation in 3G was expected, but not found, as the data revealed no significant difference between any of the test conditions. It was concluded that once the background noise calibration was performed, the system was able to effectively compensate for the aircraft noise background.

To investigate the effect of aircraft noise, breath noise and G-loads on speech recognition, a second test was performed. This time a NASA OV-10D aircraft was used, and pilots wore an oxygen mask with an Air Force standard M-169 microphone at the time of the recordings. Two speech recognition systems were evaluated: ITT VRS-1290, as before, and Verbex VAT31. The former was tested on eight subjects, and the latter one on five subjects. Recordings were made in the laboratory, on the ground with engines off, at 1G, at 4G and once more at 1G. The vocabulary contained 47 words and a language model was used. As before, the experiment began with template generation for each speaker.

The average performance over five subjects with the ITT VRS-1290 system was 97.2 % for the two ground conditions, and 92.1 % for the three flight conditions. Two factors accounted for the majority of the recognition errors: the Lombard effect and lack of automatic gain control. The average over the same five speakers for the Verbex VAT31 test was 99.5 % for the ground conditions and 97.3 % in the flight conditions.

The study concludes that robust speech interfaces in cockpits are fast becoming a cost effective control strategy, and that commercial systems developed for the automobile industry and for telephone applications can be modified to suit the fighter cockpit environment with only small adjustments.

4 The Gripen corpus

The speech data used for this project was recorded by Bengt Willén at Saab in Linköping, during the years 2000 and 2001. The purpose of the recordings was to obtain a reference material for research and development, and for the evaluation of different speech recognizers.

The recordings were made with a Sony TCD-D8 DAT-recorder (Digital Audio Tape), which had been modified to meet the requirements of the enviroment in the Gripen aircraft. The speech was picked up by one of the dynamic microphones inherent in the pilot's mask, which was connected to the DAT-recorder with an adapter made especially for this task. A sampling frequency of 48kHz was used, but the files were later downsampled to 32kHz at Saab.

4.1 The speech data

The material consists of 30 sentences, made up of words of common use in the cockpit (see appendix). The sentences range from short ones, containing just two words, to longer ones that include sequences of digits. The speakers were eight male Swedish pilots. Seated inside the aircraft, the pilots read the sentences in Swedish, and also the corresponding ones in English. This procedure was repeated in the following test conditions:

- On the ground with the engine off
- On the ground with the engine running
- During normal flight in 1G
- In 4G

In addition, some recordings were made during flight at higher G-loads, with a maximum of 8G, above which normal speech is very difficult. The sentences in these conditions were either some subset of the commands, or more commonly, a sequence of digits. Also for many, but not all speakers, there were miscellaneous files containing spontaneous speech, for instance communication with air traffic control. These recordings were made at unknown G-loads, and they made a considerable contribution to the size of the vocabularies.

The total length of the speech material is estimated to 77 minutes for Swedish, and 63 minutes for English. The difference is due to a smaller number of miscellaneous files, and of speech at high G-loads, in the English case. The files contained some silence and extra-linguistic sounds such as breathing. Approximately 20 % of the material consists of silence and breath noises.

4.2 Noise in the Gripen corpus

The Gripen corpus contains all of the distrubances and types of noises mentioned in section 3.2. Environmental noise is present and grows more evident at high G-loads. The Lombard effect is observable in some of the speakers, while others had a tendency to whisper when subjected to high G-loads. Also found in the material are more specific disturbances caused by the interaction between speaker and microphone. Examples are puffs of air, loud respiratory noises at the beginning and end of words and some occurances of signal distorsion caused by shouting at high G-loads.

A common type of noise is breath noise. From examining the spectrograms of these sounds, it can be seen that the inhalations generally contain more energy than the exhalations, and that the energy is distributed differently on the frequency scale. At high G-loads, exhalations become more noise-like, with fewer visible formants. Figures 2 and 3 show examples of the types of breathing noise common in this material.

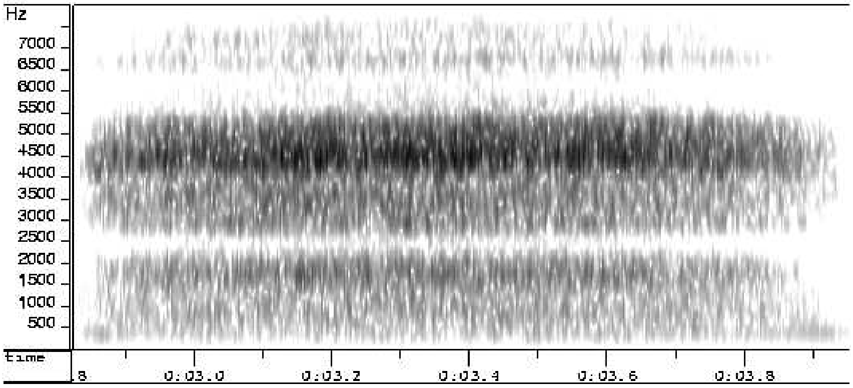

Figure 2: Spectrogram showing an example of an inhalation

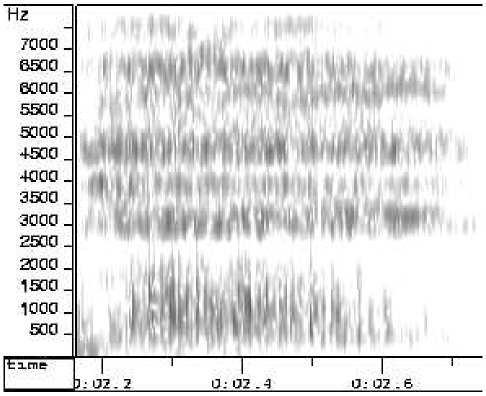

Figure 3: Spectrogram showing an example of an exhalation

5 Initial recognition tests

To give a rough idea of what results could be expected when performing recognition with the Gripen corpus, a first recognition test was performed using the entire corpus as test material. Furthermore, another test was made on the files recorded in the best condition, i.e on the ground. These files were not distorted by noise and/or strained voices, and could therefore be expected to give better results.

Throughout this project, experiments have been conducted in parallell for the two languages: Swedish and English.

5.1 Data preparation

Speech recognition is preceeded by the process of data preparation. For completeness, the parametrization of the sound files and the generation of label files are briefly described below.

5.1.1 Parametrization

The sampling frequency of the sound files was 32 kHz. Since the test material must have the same sampling frequency as the material on which the models were trained, they were downsampled to 16 kHz, using the linux sound processing program Sox. Twelve mel frequency cepstral coefficients were computed for the feature vectors, and the first and second time derivative were added, as well as the 0'th cepstral coefficient.

5.1.2 Labelling the corpus

The files were manually transcribed and labelled. This was done in accordance with the SpeeCon transcription conventions (Iskra et al., 2002).

It has been shown that explicit modelling of non-speech acoustic events improve recognition scores (Lindberg et al., 2000). Speech at high G-loads is likely to contain these sorts of sounds, and in the Gripen corpus they were quite common in the noisier conditions. According to the SpeeCon annotation standard, four non-speech labels were used: *fil*, which stands for filled pauses made by the speaker; *spk*, which represents other kinds of noise produced by the speaker, such as coughing or throat clearing; *int*, meaning intermittent noise; and finally *sta*, for stationary noise. The latter two were later ignored, and the noisy data was kept in the material, since noise must be considered normal for the environment in question.

5.2 Initial Swedish tests

The dictionaries for the two types of tests mentioned above contained 516 entries for the full database, and 61 entries for the best ground condition. An unrestricted grammar was used, in which any word could follow upon any other with the same probability.

A model set trained on the Swedish SpeeCon database was available at TMH. It consisted of 50 monophones, including short pause and silence models, as well as models for speaker noise. All phonome models had three active states and 32 Gaussian mixtures in each state. As is common, the short pause model (*sp*) shared parameters with the middle state of the silence model, so called tying.

Monophones are context independent, and can therefore be expected to perform worse than context dependent sub-word units, for instance, triphones. The reason for choosing monophones was that the models would later be used for adaptation, and the available data was insufficient to adapt context dependent models.

5.3 Initial English tests

For the English tests, a new model set was trained from scratch from the TIMIT database. A script by Giampiero Salvi at TMH was modified and used for this purpose. The training procedure was as follows:

- Initialization with the HTK tool HInit for all 49 models.[1]
- Adding and tying of the *sp* model
- Four iterations of Baum-Welch re-estimation
- Updating number of Gaussian mixtures to 32, each update followed by two passes of re-estimation

Thereafter, the same initial recognition tests were made as for the Swedish case. Dictionaries were 268 entries for the full database, and 57 entries for the best ground condition.

5.4 Reference tests

In addition, recognition tests were made with the SATSA air traffic control speech corpus, which contained clean speech in both Swedish and English. These tests were thought to be useful as an evaluation of model performance, and to be interesting when compared with the results obtained from the Gripen material. The reference tests will be described in more detail in chapter 7.

[1] Originally the TIMIT database uses a 61 phoneme symbol set, but these were mapped onto a simpler phoneme set.

6 Adaptation

As is often the case, the models used for this project were trained on clean studio speech. For reasons already mentioned, the Gripen files differed significantly from these conditions, wherefore the initial recogntion results could be expected to be rather poor. To investigate to what extent results could be improved by adaptation, a series of tests were performed. Both the Swedish and English model sets were adapted, and tested on a number of subdivisions of the Gripen corpus, as will be described below.

At this stage, models for breathing were created and used in all the adaptation procedures. Inhalations and exhalations were modeled separately. These models are hereafter refered to as *in* and *out*.

6.1 Adaptation to a large material

One of the initial recognition test used the whole corpus as test material. It was therefore interesting to make use of the whole material also in adaptation, and compare these results with the results from the initial test.

6.1.1 Definition of training and test sets

The data was divided so that one speaker at a time acted as test speaker, while the data from the remaining speakers was used for adaptation. This procedure was repeated for each speaker and for both languages. The idea was that the models would be adapted to the environment and the channel in a general way, and therefore improve results for the speaker whose data was not present in the training material.

6.1.2 Adaptation procedure

The adaptation was carried out in a re-training procedure as follows:

- Flat start initiation of models *in* and *out* [2]
- Three passes of re-estimation using all models
- Alignment and detection of outliers followed by one iteration of re-estimation
- Successive updating of the number of mixtures to 32 for models *in* and *out*, each updating followed by two iterations of re-estimation
- Two more iterations[3] of re-estimation using the 32 mixture *in* and *out* models and the original phoneme models.[4]

The English model set originally lacked models for speaker noise. To make results comparable with the Swedish case, the *spk* and *fil* models were taken from the Swedish model set and used in re-traing. This was possible since both model sets had the same topology and properties.

[2] The means and variances of all states set to the global values of the training set.

[3] The optimal number of iterations was found experimentally.

[4] The models already subjected to extensive re-estimation were discarded to avoid overtraining.

6.2 Adaptation to specific test conditions

Adaptation was also performed with respect to each unique test condition (see section 4.1). For these test, however, a re-training procedure as the one described above was not feasible because of data sparsity. Instead, MLLR together with MAP was used (see section 2.3), since these techniques have been shown to perform better when used together, compared to each one in isolation (Young, 2002).

6.2.1 Definition of training and test sets

The following test conditions each formed an adaptation category:

- Speech in the ground condition with the engine off
- Speech in the ground condition with the engine running
- Speech in 1G
- Speech in 4G
- Utterances in 6, 7 and 8G were grouped togheter[5]

The size of the dictionaries were between 50 and 60 entries, and the grammar was unrestricted. The same round robin procedure as in the previous test was used, i.e. the test speaker was alternated and the remaining 7 were used for training.

6.2.2 Adaptation procedure

For both languages, the phoneme models as well as models for breathing and for speaker noise were adapted in the following procedure:

- The generation of a regression tree with 16 base classes
- Global MLLR adaptation with HEAdapt
- A second adaptation pass with HEAdapt, using MLLR and MAP with a scaling factor of 15.0[6]

Thereafter, the test utterances in each category were tested on both the unadapted and the adapted models.

[5] There was not enough data to adapt these conditions separately, since speech at higher G-loads was not present in the data for some of the speakers

[6] The optimal scaling factor was found experimentally

7 SATSA

As an extension to this thesis, experiments were made with the SATSA database. The work included data preparation and recognition tests. The tests were thought to be a useful reference; an evaluation of model performance on clean speech.

SATSA (Swedish Air Traffic Services) is a training school for air traffic controllers. The SATSA database was recorded in 2001 to be used for recognition experiments, the long-term goal being to make speech technology a part of the flight simulator training.

A total of thirty Swedish and seventy English sentences, spoken by seventeen Swedish speakers, constitute the SATSA database. Six of the speakers were women. The sentences were strings of digits mixed with flight control vocabulary.

7.1 Data preparation

The sound files had a 16 kHz sampling frequency, which meant that they did not need to be downsampled. As with the Gripen files, MFCCs were computed and the first and second time derivative was added, as well as the 0'th cepstral coefficient.

Transcriptions existed, but were not labelled for deviations and non-speech events. Going through each individual file would have been too time consuming. Therefore, a few random samples were taken for each speaker in both languages, and as these turned out to be correct, the remaining files were assumed to be reasonably correct.

The sizes of the Swedish and English dictionaries were 101 and 96 words respectively, and the grammar was unrestricted.

7.2 Tests

The material was divided into Swedish and English, and recognition tests were performed on the whole material, using the same models as before. In the English case, no *spk* and *fil* models were used, since these labels were not used in the transcriptions.

8 Results

This section presents the results obtained from the the recognition tests and the adaptation performed in this work. After explaining how results should be interpreted, the results from the initial tests are presented, followed by the results from the various adaptation procedures. Detailed data is given in the appendix.

8.1 Word Accuracy

The evaluation of the results was done with tools from the HTK toolkit. The output of the recognizer and the true transcriptions were compared, and the optimal string match found.

Two ways of presenting the results are common. The number of correctly recognized labels is given by:

$$\%Correct = \frac{N - D - S}{N} * 100 = \frac{H}{N} * 100 \tag{1}$$

Whereas the word accuracy is defined as:

$$Accuracy = \frac{N - D - S - I}{N} * 100 = \frac{H - I}{N} * 100 \tag{2}$$

Where,

- H - number of correct labels
- D - number of deletions
- S - number of substitutions
- I - number of insertions
- N - total number of labels

8.2 Word insertion penalty

The word insertion penalty is a value that is added to the score at each transition between two words. When the word insertion penalty is low, the recognizer can insert words without much extra cost. If there is a high number of insertions, the word accuracy will decrease - it may even be negative - even though the percentage of correctly recognized words still may be high. These two measures can therefore give rather different results. This was striking upon inspection of the recognition results for the Gripen material; there was no insertion penalty value that optimized both these measures simultaneously, but instead the % correct words could drop substantially before the accuracy reached a peak and vice versa.

However, word accuracy can be said to be the more representative figure of recognizer performance (Young, 2002). In the following, results will therefore be presented as word accuracies. Both measures are given in the appendix.

Furthermore, it should be mentioned that the results presented in this section have been optimized for each test, and need not have the same word insertion penalty.

8.3 Initial tests

In the first of the two initial tests, all utterances were used as test material. The second test used only files in the ground condition with the engine off. Once again, the Swedish and English dictionaries were 516 and 268 entries for the former case, and 61 and 57 entries for the ground condition.

As can be observed in figure 4, the word accuracies for Swedish and English were 9.79 % and 13.23 % respectively. The higher bars show the results from the second test, which used the speech recorded in the ground condition. These scores were 46.73 % for Swedish and 32.28 % for English.

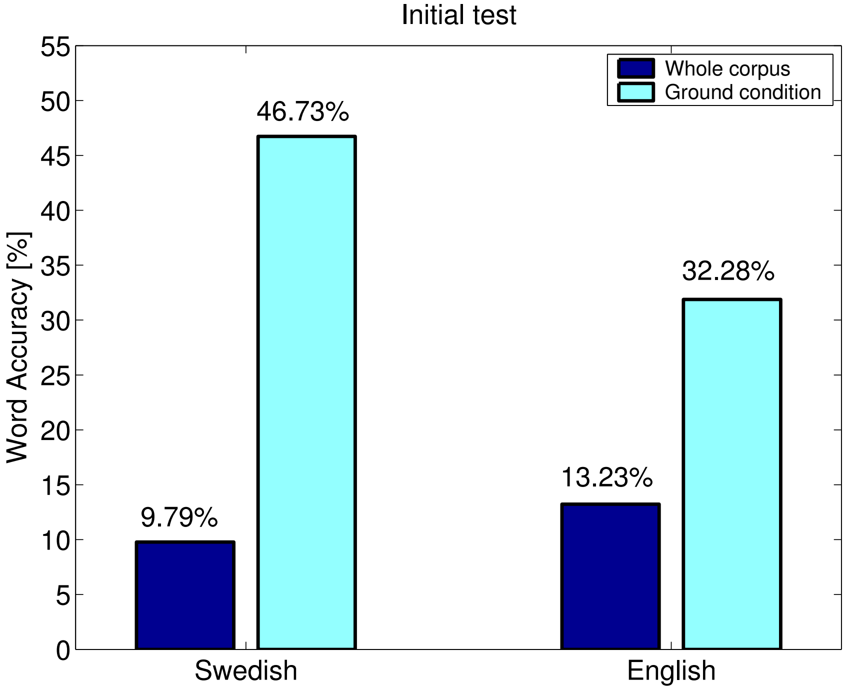

Figure 4: Results for Initial recognition tests

8.4 SATSA

As was expected, the clean speech of the SATSA corpus recieved higher scores than the Gripen corpus. Swedish was recognized slightly better, having a word accuracy of 98.70 %, compared to 97.10 % in the English case. Both dictionaries contained approximately 100 words.

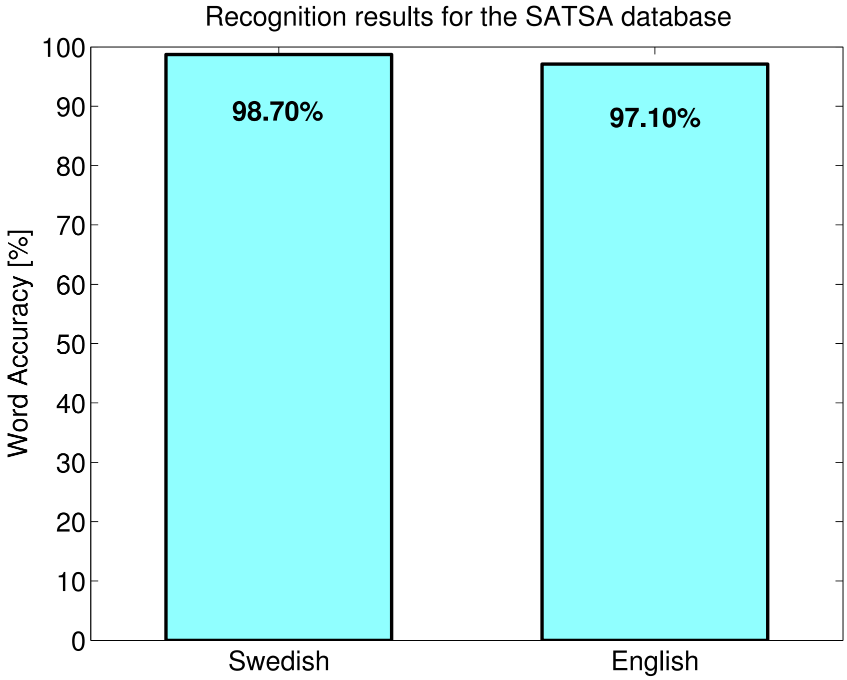

Figure 5: Recognition results for SATSA.

8.5 Adaptation to a large material

In re-training on a large material, the individual speaker results and the means are presented in the table below, and in figures 6 and 7. All speakers except the test speaker are used for re-training.

A dramatic improvement is seen when comparing these results to the initial test. Here the means for the word accuracy was approximately 59 % for Swedish, and 63 % in the English case.

Speakers	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**Mean**
Swedish	46.58	45.53	55.21	61.56	65.95	69.37	64.93	67.13	**59.53%**
English	68.26	65.85	63.98	71.73	58.28	61.12	51.88	65.41	**63.31%**

Table 1: Results for adaptation to a large material.

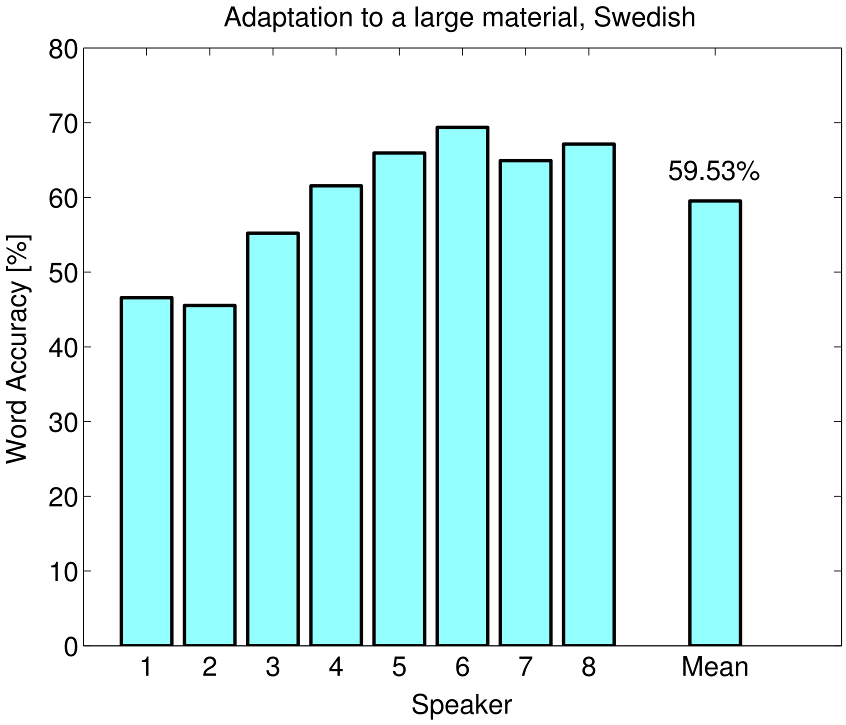

Figure 6: Results for adaptation to a large material, Swedish.

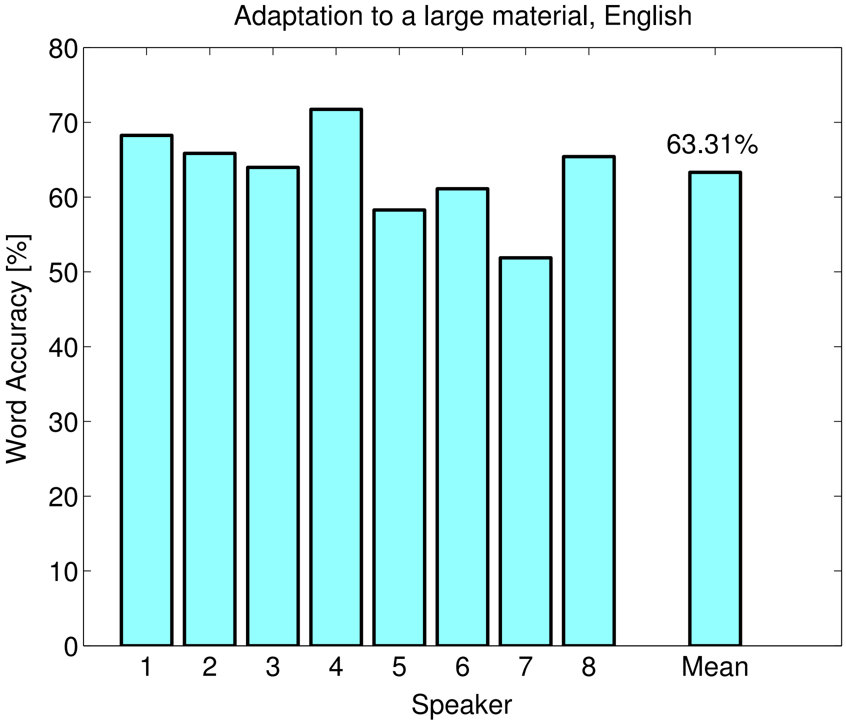

Figure 7: Results for adaptation to a large material, English.

8.6 Adaptation to specific test conditions

Adaptation results for the two ground conditions (with and without the engine running), 1G, 4G and higher G-loads are presented below and in that order. All speakers except the test speaker are used for re-training.

8.6.1 Ground condition, engine off

The unadapted and adapted results for each speaker, as well as the means, can be found in the tables below. Figures 8 and 9 give the same results in bar charts; the darker bars being the unadapted word accuracies, and the total height of each bar representing the adapted results.

As can be observed, the unadapted means of this relatively clean speech are both close to 60 %, the Swedish figure being higher by 4 percentage points. After adaptation, both results are close to eighty %, the Swedish results remaining better by approximately the same amount.

The improvement upon adaptation were 33 % in the Swedish case, and 34 % for English.

Speakers	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**Mean**
Unadapted	64.61	58.62	44.44	74.65	62.94	61.81	74.29	60.13	**62.69%**
Adapted	77.53	80.46	75.56	85.21	94.42	85.42	79.05	88.61	**83.28%**

Table 2: Results for Swedish adaptation to the ground condition.

Speakers	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**Mean**
Unadapted	54.59	57.98	47.37	71.68	52.69	57.74	78.18	49.11	**58.67%**
Adapted	72.97	77.66	77.37	87.28	67.07	76.79	88.18	82.25	**78.70%**

Table 3: Results for English adaptation to the ground condition.

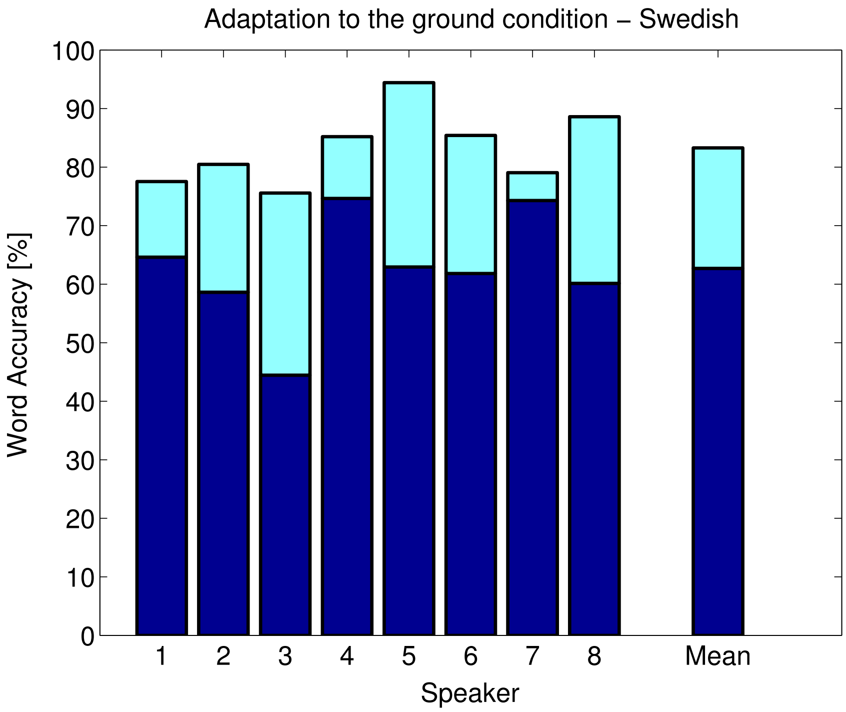

Figure 8: Adaptation to the ground condition, Swedish.

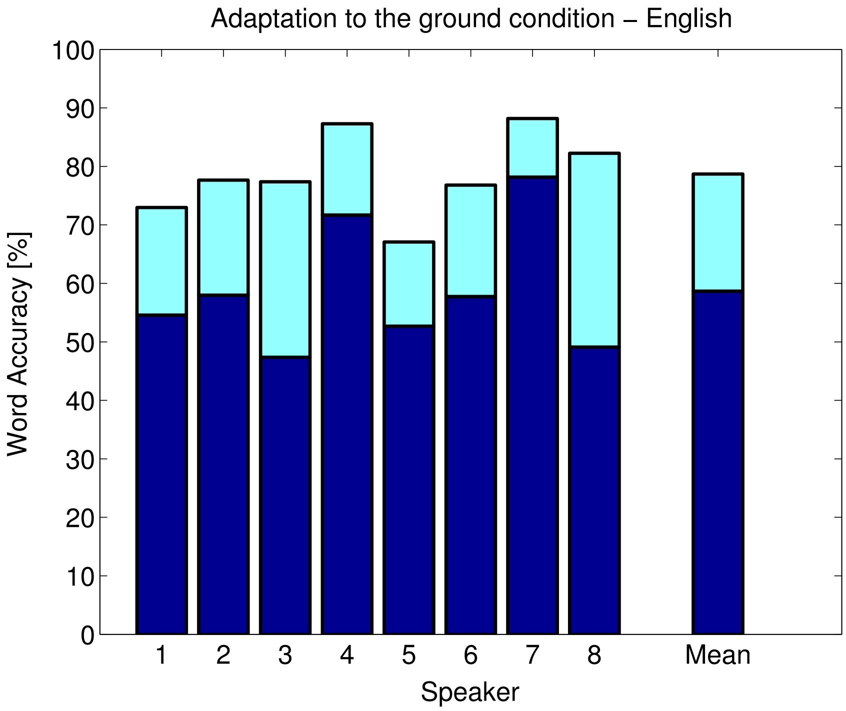

Figure 9: Adaptation to the ground condition, English.

8.6.2 Second ground condition, engine on

Figures 10 and 11, as well as tables 4 and 5 show the results before and after adaptation to the ground condition with the engine on.

Also here the results after adaptation are considerably better. The improvement was 86 % and 53 % for Swedish and English. The adapted results for the two languages differ by less than two percentage points, even though there was a 10 percentage points difference initially.

Speakers	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**Mean**
Unadapted	43.62	34.67	37.27	47.26	50.52	40.43	45.83	42.67	**42.78%**
Adapted	71.14	69.33	84.47	76.03	80.93	82.98	83.33	84.00	**79.03%**

Table 4: Results for Swedish adaptation to the ground condition with engine switched on.

Speakers	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**Mean**
Unadapted	58.79	37.35	38.65	62.03	57.75	45.58	72.22	50.94	**52.91%**
Adapted	81.87	75.30	74.85	77.85	82.89	80.27	87.04	85.53	**80.70%**

Table 5: Results for English adaptation to the ground condition with engine switched on.

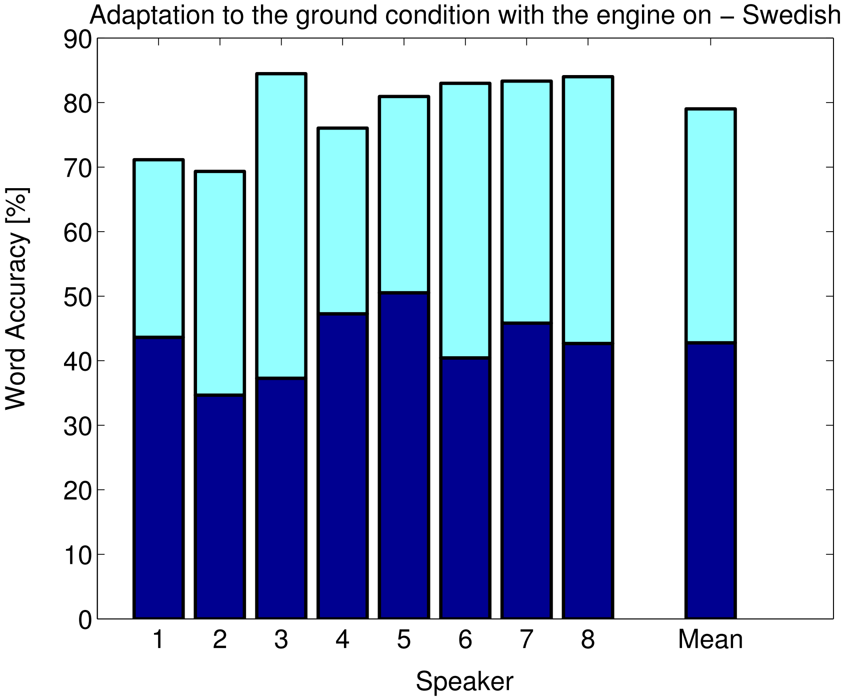

Figure 10: Adaptation to the ground condition, engine on, Swedish.

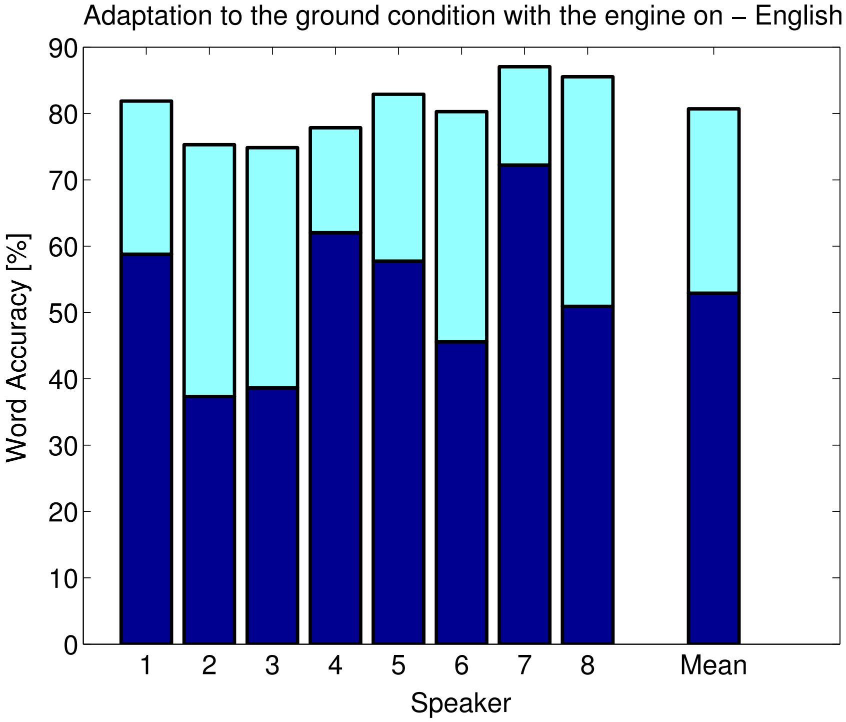

Figure 11: Adaptation to the ground condition, engine on, English.

8.6.3 1G

Results after adaptation in 1G were approximately 73 % and 66 % for Swedish and English respectively, compared to 36 and 37 % before adaptation. In the former case, adaptation improved the word accuracy by almost 99 %, whereas the English results were 75 % better after adaptation.

Speakers	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**Mean**
Unadapted	36.54	35.03	31.49	42.58	47.69	29.85	39.42	31.88	**36.33%**
Adapted	75.00	67.80	64.64	68.39	82.50	79.10	76.92	71.25	**73.14%**

Table 6: Results for Swedish adaptation to speech in 1G.

Speakers	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**Mean**
Unadapted	42.29	23.46	19.23	55.25	35.42	45.03	50.49	28.90	**37.50%**
Adapted	67.40	65.43	56.59	72.55	60.42	80.79	72.82	54.91	**66.30%**

Table 7: Results for English adaptation to speech in 1G.

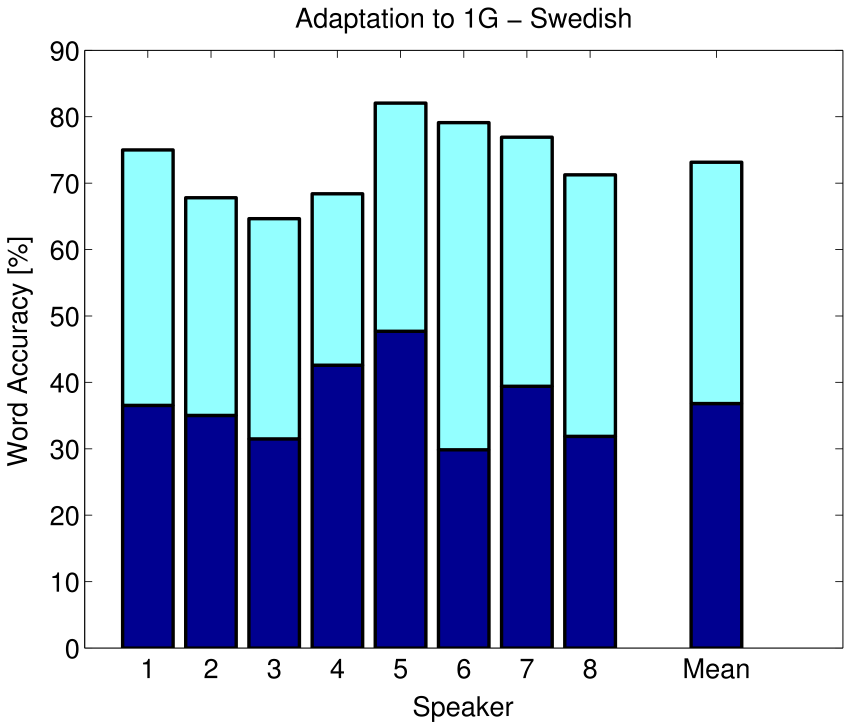

Figure 12: Adaptation to speech in 1G, Swedish.

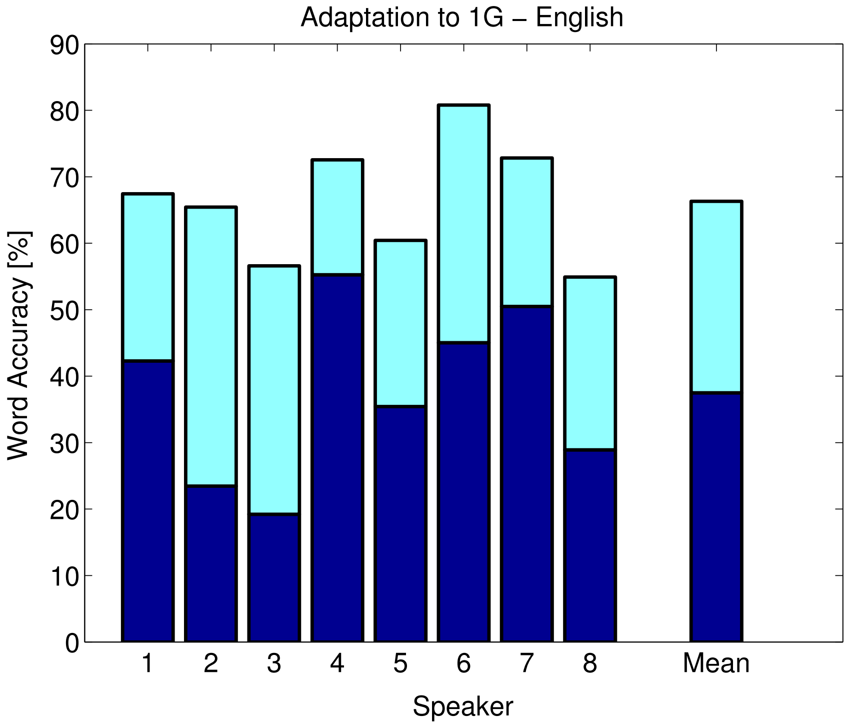

Figure 13: Adaptation to speech in 1G, English.

8.6.4 4G

The Swedish files were more correctly recognized both before and after adaptation. The means for all speakers rose from approximately 40 % (Swedish) and 30 % (English), to 72 and 59 %.

Improvements were 79 and 92 %.

Speakers	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**Mean**
Unadapted	45.67	35.80	30.87	51.63	47.06	39.86	35.42	34.95	**40.16%**
Adapted	72.12	59.88	67.11	76.09	79.19	78.38	75.00	66.67	**71.80%**

Table 8: Results for Swedish adaptation 4G.

Speakers	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**Mean**
Unadapted	41.8	11.11	19.19	42.08	28.17	38.69	40.91	22.40	**30.54**
Adapted	61.38	49.74	61.63	64.48	52.58	66.07	64.55	48.63	**58.63**

Table 9: Results for English adaptation 4G.

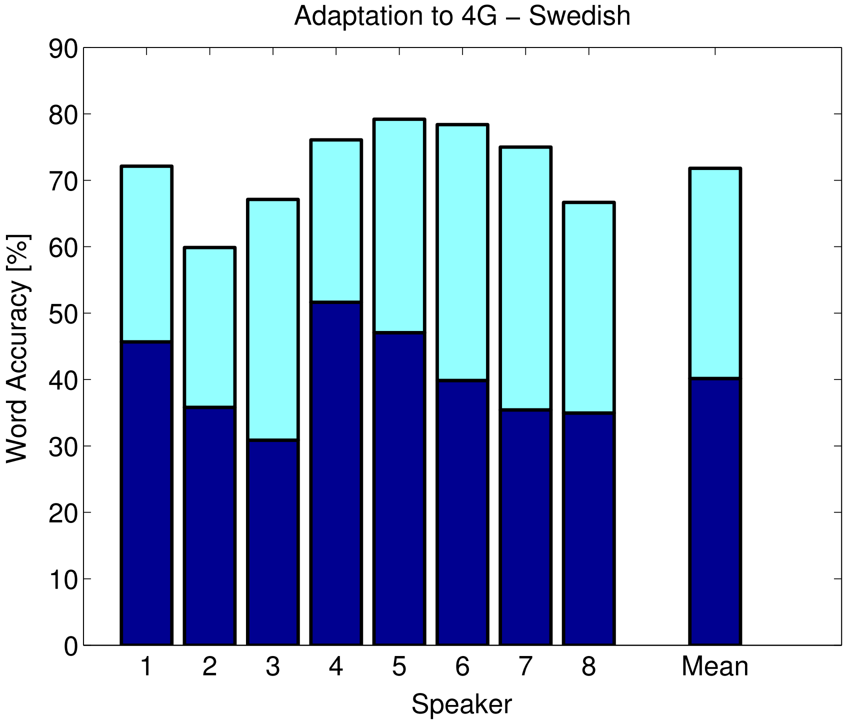

Figure 14: Adaptation to speech in 4G, Swedish.

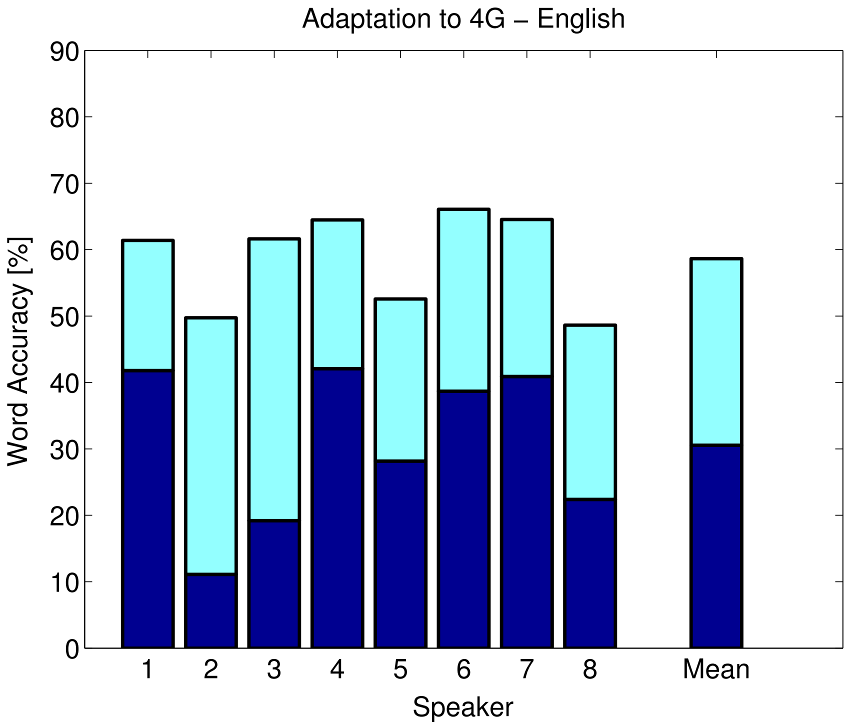

Figure 15: Adaptation to speech in 4G, English.

8.6.5 Higher G-loads

All G-loads higher than 4G were grouped togheter. As mentioned, adaptation was performed using speech at 6 and 8G, and the test material was speech at 7G. The results are presented in table 10. The bar chart gives the same information; the unadapted test results for both languages as dark bars. These were close to 38 % and 51 % respectively. The word accuarcies after adaptation are given by the total heights of the bars. They were 50 % and 59 %.

A 32 % improvement was acchieved for Swedish. The same figure for English was 14 %. This is strikingly lower than for the other test conditions.

	Unadapted	**Adapted**
Swedish	38.06	50.32
English	51.32	58.55

Table 10: Results for G-loads 6, 7 and 8.

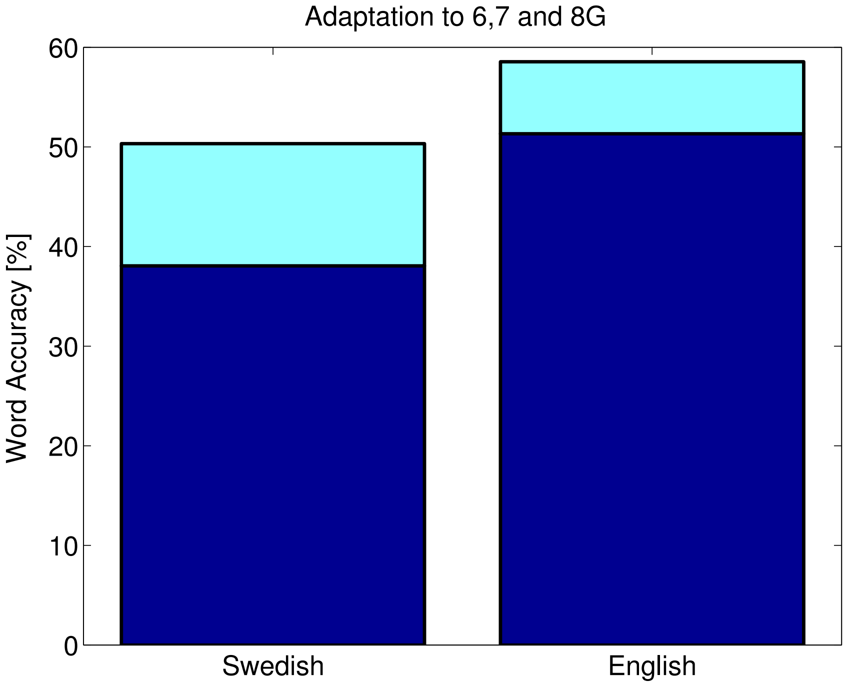

Figure 16: Results for adaptation 6, 7 and 8G.

9 Discussion

This section will focus on discussing the results presented in the previous section. A summary and conclusion follow in the next chapter.

First some general comments will be made about the results. The reference test, i.e. recognition on the SATSA material, showed that the models used in the experiments were able to give a good performance. A small degradation in performance due to the labels not being thoroughly checked for errors can not be excluded. Nevertheless, the models performed well, also for English, although the speakers were Swedes and therefore have an accent quite different from the native American speakers used for training.

Also noteworthy is the fact that the models were all monophones, which are known to perform less well than context dependent units. The reason for using monophones was the relatively small amount of data that was available for these experiments. Adaptation is only as successful as the training data permits, something which should be kept in mind when interpreting these results. Some of the tests indicate that better results would have been achieved had the amount of data been larger.

There is a significant variation in articulation between speakers at high G-loads, due to individual G-tolerance and speaking style. Nevertheless, dividing the material according to G-loads, and not on a speaker basis, was the only feasible approach given the amount of available data.

9.1 Initial tests

The two initial tests confirmed the expectation that the Gripen material could be difficult to recognize. For the one using the entire corpus, the word accuracies were in both cases below 15 %, which must be seen as extremely low, also in comparison with the SATSA corpus. Higher scores can be observed for the test that used only the fairly clean speech of the ground condition. These figures were just above 45 and 30 %, which is clearly better, but still low compared to most real life recognition tasks. Factors that are likely to have complicated recognition over all are the environment and recording channel, for reasons already mentioned, and above all, the prominent and frequent breathing sounds for which there were no models in these tests. In many cases, these caused the recognizer to assign one of the words in the dictionary to these accoustic events, something which reduces the score and also might effect the recognition of the surrounding words.

In comparing the Swedish and English results, a few comments can be made. In testing the whole material, the English corpus had two advantages, one being a dictionary half the size of the Swedish one, and the other being that there were fewer English files containing spontaneous speech. The smaller dictionary is advantageous in that there are fewer words that can be substituted for one another, while spontaneous speech can be problematic in the sense that it is full of filled pauses, silence, hesitation and so on, which poses a challenge to most recognizers. The higher scores for English in the first initial test is probably explained by the last fact, and possibly also by the size of the dictionary. For the ground condition, the Swedish advantage is harder to explain. It is likely that the broken English of the speakers played a role here. However it should not be the sole cause for this rather large difference. One hypothesis is that the small amount of data in this test caused individual files to have a large impact on the results.

9.2 Adaptation

It was hoped that performing adaptation to a large part of the corpus would improve results by adapting the models to the sound environment in general. Better results could also be expected since the effect of the channel should be reduced after adaptation. The models should also become more familiar with the broken English of the speakers. Finally, as the models for breathing were included in the adaptation, they should hopefully contribute to higher scores.

Indeed, the recognition tests that followed the adaptation show improvements in the word accuracy for both languages, which were in both cases around 60 %, a figure which is quite high compared to the initial test that made use of the whole material (9 and 13 %).

It is mentioned in section 4.1 that there was a larger amount of data in the Swedish part of the corpus. For this reason it might be expected that the Swedish results would benefit from this and exceed the English ones. Upon inspection of the results, this was not the case, as the English mean was in fact about four percentage points better. One possible reason is that the additional material consisted of spontaneous speech, and of course, these files were all present in the test set at some point (since the re-training used a round robin procedure). Another reason why the spontaneous speech files recieved a low score is that they contained words that were not part of the intended vocabulary, and therefore were not present elsewhere in the material. This would mean that there are likely to have been combinations of monophones that is not guaranteed to have been encountered by the recognizer during training.

Even though a distinct improvement is observed after adaptation to a large material, the approach is not optimal. To see if further improvement could be achieved, adaptation was carried out with respect to the individual test conditions (G-loads).

Figure 17 plots the results before and after adaptation for the two ground conditions (with and without the engine running) and for 1G and 4 Gs. The test conditions are found on the x-axes and word accuracy on the y-axes. The dashed lines are the means of the adapted results and the solid lines display the means before adaptation. Squares are used to represent Swedish and triangles are used for English.

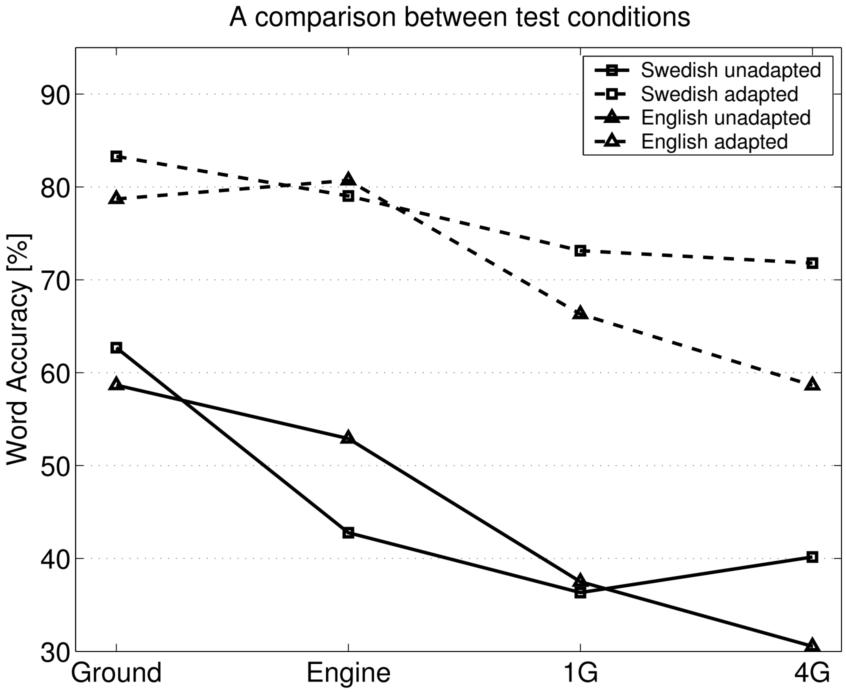

Figure 17: The effect of G-loads on recognition.

The most important tendency that can be seen in the figure is the decreasing accuracy of the unadapted results. The test and adaptation materials, as well as the dictionaries were the same sizes in these four tests, for which reason this comparison should be completely valid. In other words, it is found that increasing G-loads complicate speech recognition.

As is also observable in the figure, the adapted results keep well above the unadapted ones in all cases. The tendency is for high G-loads to result in lower scores even after adaptation, which is quite natural given that the models already from the start had shown to be less capable of correctly recognizing this speech. More adaptation data would be required for the noisier conditions, to obtain good adaptation results.

It is also worth pointing out that the scores after adaptation were better then that of adaptation on the whole material, the only exception is the result for adaptation to English in 4 Gs, which was worse by a small amount. The G-load specific tests used a smaller dictionary, and only a small fraction of the training data in the first test. But even so, the tendency seems to suggests that it is a good idea to be more specific when adapting.

A descrepancy can be seen for the unadapted Swedish result in 4G; the score is higher than that for the 1G condition. There are no obvious causes for this deviation, and it is likely that a larger amount of data would have produced different results, more in line with what could be expected.

The results for higher G-loads are not included in the figure since this test is not comparable to the others. It involved adaptation to serveral G-loads simultaneously, and the division of the material was not done on a speaker basis. Furthermore, the amount of data was smaller since these files contained mostly digits.

9.3 Higher G-loads

The files recorded at 6, 7 and 8G were by far the noisiest ones, some of them difficult to make out even for a human being. Adaptation to these conditions was done with just a small amount of data, and was the least successful one in terms of improvement of the scores.

There was quite a large difference in results for the two languages, English performing better before adaptation than Swedish after the same. There is no obvious reason for this, but it should be kept in mind that the content, as well as the articulation, at these high G-load files varied a lot. Therefore, it seems unreliable to draw conclusions from these results.

9.4 The effect of introducing models for breathing

To see whether modeling the breath noise had a positive effect on recognition, two tests were made on speech in the ground condition. One of the tests included models for breathing, the other did not.

Figure 18 shows the results side by side. A substantial increase in the word accuracies for both languages is observable, which indicates that the breath models helped in recognizing these files.

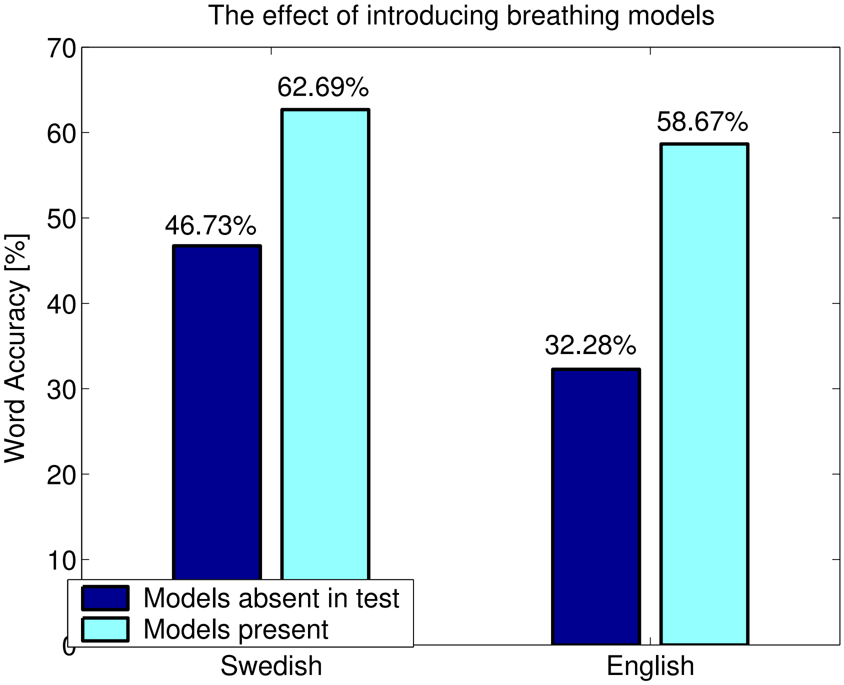

Figure 18: The effect of introducing breathing models.

9.5 A comparison between languages

Concerning the recognition results for the English material, it could be expected that the accent of the speakers would have a negative effect on the test results, since the original model set was trained using native American English speakers.

Earlier in this chapter, the results for the initial recognition tests were discussed. It was found there that the English material recieved higher scores than Swedish when the entire corpus was used, and the probable reasons were discussed. When testing only the files of the ground condition, English scores were significantely lower than Swedish, but it is unclear to what extent this was caused by the accent of the speakers.

When looking at the results of the recognition tests for the individual test conditions (see figure 17), the lines showing the Swedish and English results before adaptation intersect on two occasions, meaning that there was not any tendency for either language to recieve higher scores.

It is important to know that the tests carried out here were not designed with this question in focus, so other variables may account just as well for the results, such as dictionary sizes and an uneven composition of the data for the two languages, and above all, the two model sets were trained from different databases.

9.6 Individual speaker performance

Figures 19 and 20 show the individual speaker results for recognition with the original model sets; comparing speaker results after adaptation can be misleading, since the results would depend on the adaptation data, i.e on several speakers other than the one of interest.

For each speaker, the four bars represent the results for the four test conditions, and the horisontal lines are the mean values for the test conditions across all speakers.

Some differences between the speakers can be observed from these graphs. Speakers 2,3 and 8 were consistently below the mean for all test conditions in both languages. Speaker 4 was the only speaker to be above the mean for both languages.

No conclusions about individual sensitivity to G-loads can be drawn from this material. There was no data other than for 1G and 4G, and only two of the speakers performed worse in 4G for both languages.

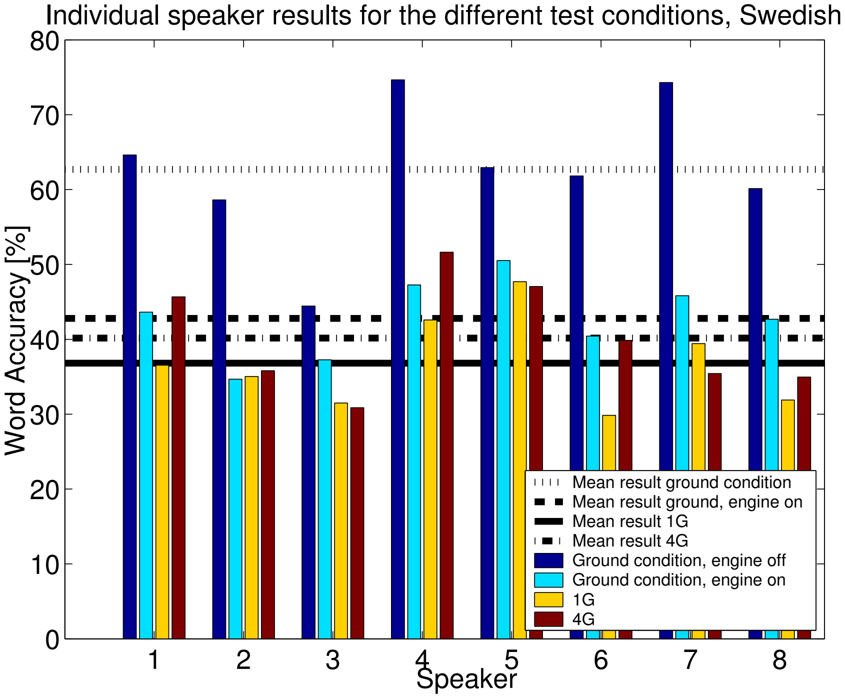

Figure 19: Results for individual speakers, Swedish

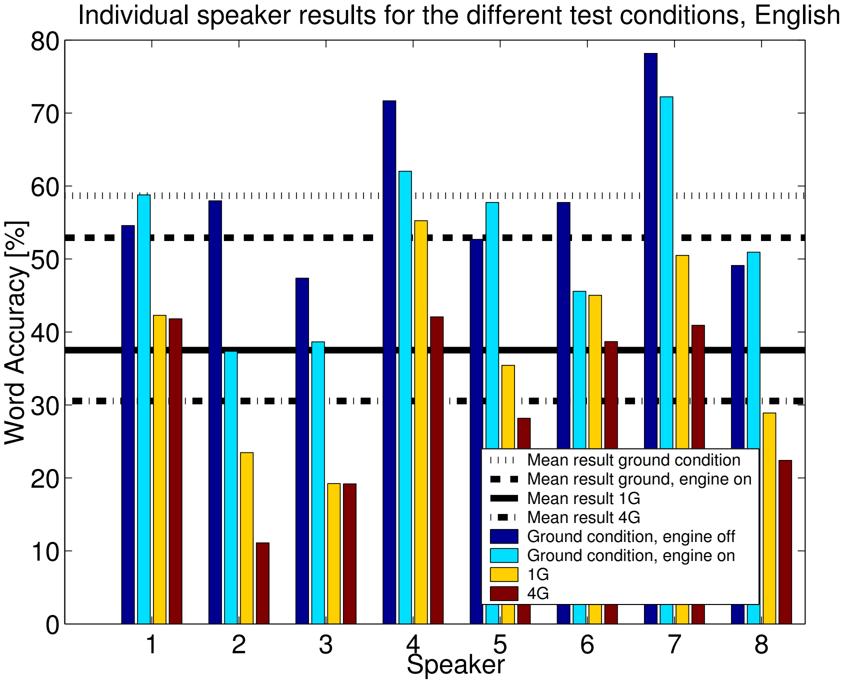

Figure 20: Results for individual speakers, English

10 Conclusion

It was found in this thesis that recognition deteriorates with increasing G-loads. It can also be concluded that adaptation greatly improved the results in all cases.

Furthermore, introducing models for breathing proved to be a good approach; these were shown to improve recognition scores significantly.

Contrary to what might be expected, no effect of the broken English of the speakers were found. Other factors had larger impact on the recognition results.

It was evident that the spontaneous speech caused problems for the recognizer, as could be expected. Any real application should therefore include a restricted vocabulary, and above all, a proper syntax, which can be expected to improve recognition accuracy substantially.

10.1 Outlook

Even though the scores were greatly improved upon adaptation, they are still too low to be acceptable for any real system. It would be interesting to repeat these experiments, including adaptation, after having applied some form of noise compensation.

Other tests on the same material should include using a syntax, and if more data had been available, it would have been interesting to perform speaker adaptation, and also to see what triphone models could have added to the recognition accuracy.

References

[1] M. Blomberg and K. Elenius. Automatisk igenkänning av tal. Institutionen för tal, musik och hörsel, KTH, 2003.

[2] D. F. Gordon. Voice recognition and systems activation for aircrew and weapon system interaction. In *Aerospace and Electronics Conference, 1990. NAECON 1990.,Proceedings of the IEEE 1990 National*, 1990.

[3] K. Hammar. Möjligheten och behovet av att införa automatiska taligenkänningssystem i stridsflygplan. Master's thesis, KTH, 1995.

[4] J. Holmes and W. Holmes. *Speech Recognition and Synthesis.* Taylor and Francis, 2001.

[5] X. Huang, A. Acero, and H.-W. Hon. *Spoken Language Processing - A Guide to Theory, Algoritm, and System Development.* Prentice Hall, 2001.

[6] D. Iskra, B. Grosskopf, K. Marasek, H. van den Huevel, F. Diehl, and A. Kiessling. Speecon - speech databases for consumer devices: Database specification and validation. In *Proceedings LREC 2002, LAS PALMAS, CANARY ISLANDS - SPAIN*, 2002.

[7] B. Lindberg et al. A noise robust multilingual reference recognizer based on speechdat(ii). In *Proc. ICSLP, International Conference on Spoken Language Processing*, 2000.

[8] J.-P. Martens et al. Final report of cost action 249 - continuous speech over the telephone. Technical report, Electronics and Information Systems, Ghent University, May 2000.

[9] A. V. Oppenheim and A. S. Willsky. *Signals and Systems.* Prentice-Hall, 1997.

[10] L. Rabiner and B. Juang. An introduction to hidden markov models. *IEEE ASSP Magazine*, 1986.

[11] O. Siohan, C. Chesta, and C.-H. Lee. Joint maximum a posteriori estimation of transformation and hidden markov model parameters. *IEEE Transactions on Speech and Audio Processing*, May 2001.

[12] B. J. Stanton, L. H. Jamieson, and G. D. Allen. Acoustic-phonetic analysis of loud and lombard speech in simulated cockpit conditions. In *Proceedings, ICASSP 88, New York, NY*, April 1988.

[13] H. J. M. Steeneken et al. Potentials of speech and language technology systems for military use: an application and technology oriented survey. Technical report, NATO DRG Document AC/243 (Panel 3) TR/21, 1996.

[14] V. Vasiletz and O. Yakimenko. The concept of on-board diagnostics, prognosis and correction of pilot condition under the action of high level g-load complex. In *Aerospace and Electronics Conference, 1995. NAECON 1995.,Proceedings of the IEEE 1995 National, Volume:1*, May 1995.

[15] D. T. Williamson. Robust speech recognition interface to the electronic crewmember: Progress and challenges. In *Proceedings of 4th Human-Electronic Crewmember Workshop, Kreuth, Germany.*, September 1997.

[16] D. T. Williamson, T. P. Barry, and K. K. Ligget. Flight test results of itt vrs-1290 in nasa ov-10. In *Proceedings of AVIOS '96 15th Annual International Voice Technologies Applications Conference, (pp. 33-40), San Jose, CA: American Voice Input/Output Society*, July 1996.

[17] S. Young et al. *The HTK book, version 3.2*, 2002.

Appendix

A Recognition results for the clean SATSA database

	Swedish	English
Number of words	11067	15265
Number of correctly recognized words	10939	14866
Deletions	64	91
Substitutions	64	308
Insertions	16	43

	Word Accurcy	% Correct
Swedish	98.70%	98.84%
English	97.10%	97.39%

B Recognition results for the initial recognition tests on the Gripen material

B.1 General data

	Word Accurcy	% Correct
Swedish whole corpus	9.79%	12.38%
Swedish ground condition	46.73%	52.24%

	Word Accurcy	% Correct
English whole corpus	13.23%	16.90%
English ground condition	32.28%	54.95%

B.2 Recognition results for the whole corpus

	Swedish	English
Number of words	5864	5261
Number of correctly recognized words	726	889
Deletions	1642	1202
Substitutions	3496	3170
Insertions	152	195

B.3 Recognition results for the ground condition

	Swedish	English
Number of words	980	1010
Number of correctly recognized words	512	555
Deletions	156	61
Substitutions	312	394
Insertions	54	229

C Recognition results for adaptation to a large material

C.1 Swedish

Speaker	Word Accurcy	% Correct
1	46.58%	50.95%
2	45.53%	49.51%
3	55.21%	60.14%
4	61.56%	67.30%
5	65.95%	70.91%
6	69.37%	77.66%
7	64.93%	73.28%
8	67.13%	71.10%
Mean	59.53%	65.10%

Speaker	1	2	3	4	5	6	7	8
Number of words	1215	1028	710	627	1028	591	479	654
Number of correctly recognized words	619	509	427	422	729	459	351	465
Deletions	228	195	61	60	114	16	31	51
Substitutions	368	324	222	145	185	116	97	138
Insertions	53	41	35	36	51	49	40	26

C.2 English

Speaker	Word Accurcy	% Correct
1	68.26%	72.64%
2	65.85%	69.53%
3	63.98%	67.74%
4	71.73%	77.07%
5	58.28%	65.78%
6	61.12%	65.92%
7	51.88%	53.92%
8	65.41%	68.21%
Mean	63.31%	67.60%

Speaker	1	2	3	4	5	6	7	8
Number of words	731	896	744	750	1172	895	931	714
Number of correctly recognized words	531	623	504	578	771	590	502	487
Deletions	65	96	47	29	150	71	148	49
Substitutions	135	177	193	143	251	234	281	178
Insertions	32	33	28	40	88	43	19	20

D Recognition results for adaptation to specific test conditions

D.1 The ground condition, engine off

D.1.1 Swedish

Recognition results before adaptation:

Speaker	Word Accurcy	% Correct
1	64.61%	70.79%
2	58.62%	62.07%
3	44.44%	46.22%
4	74.65%	83.80%
5	62.94%	63.96%
6	61.81%	67.36%
7	74.29%	80.95%
8	60.13%	63.92%
Mean	62.69%	67.38%

Speaker	1	2	3	4	5	6	7	8
Number of words	178	174	225	142	197	144	105	158
Number of correctly recognized words	126	108	104	119	126	97	85	101
Deletions	7	29	72	7	19	12	3	13
Substitutions	45	37	49	16	52	35	17	44
Insertions	11	6	4	13	2	8	7	6

Recognition results after adaptation:

Speaker	Word Accurcy	% Correct
1	77.53%	89.89%
2	80.46%	89.08%
3	75.56%	83.11%
4	85.21%	91.55%
5	94.42%	99.49%
6	85.42%	88.89%
7	79.05%	84.76%
8	88.61%	89.87%
Mean	83.28%	89.67%

Speaker	1	2	3	4	5	6	7	8
Number of words	178	174	225	142	197	144	105	158
Number of correctly recognized words	160	155	187	130	196	128	89	142
Deletions	4	3	7	3	0	6	6	4
Substitutions	14	16	31	9	1	10	10	12
Insertions	22	15	17	9	10	5	6	2

D.1.2 English

Recognition results before adaptation:

Speaker	Word Accurcy	% Correct
1	54.59%	63.24%
2	57.98%	63.83%
3	47.37%	50.53%
4	71.68%	75.14%
5	52.69%	67.07%
6	57.74%	67.26%
7	78.18%	80.00%
8	49.11%	62.13%
Mean	58.67%	66.15%

Speaker	1	2	3	4	5	6	7	8
Number of words	185	188	190	173	167	168	110	169
Number of correctly recognized words	117	120	96	130	112	113	88	105
Deletions	20	16	47	16	12	17	5	9
Substitutions	48	52	47	27	43	38	17	55
Insertions	16	11	6	6	24	16	2	22

Recognition results after adaptation:

Speaker	Word Accurcy	% Correct
1	72.97%	77.84%
2	77.66%	84.57%
3	77.37%	80.00%
4	87.28%	89.02%
5	67.07%	88.02%
6	76.79%	85.12%
7	88.18%	89.09%
8	82.25%	82.84%
Mean	78.70%	84.66%

Speaker	1	2	3	4	5	6	7	8
Number of words	185	188	190	173	167	168	110	169
Number of correctly recognized words	144	159	152	154	147	143	98	140
Deletions	12	3	10	7	3	5	2	8
Substitutions	29	26	28	12	17	20	10	21
Insertions	9	13	5	3	35	14	1	1

D.2 The ground condition, engine on

D.2.1 Swedish

Recognition results before adaptation:

Speaker	Word Accurcy	% Correct
1	43.62%	55.70%
2	34.67%	39.33%
3	37.27%	40.99%
4	47.26%	62.33%
5	50.52%	54.64%
6	40.43%	51.77%
7	45.83%	56.25%
8	42.67%	49.33%
Mean	42.78%	51.30%

Speaker	1	2	3	4	5	6	7	8
Number of words	149	150	161	146	194	141	96	150
Number of correctly recognized words	83	59	66	91	106	73	54	74
Deletions	20	29	46	10	33	10	20	29
Substitutions	46	62	49	45	55	58	22	47
Insertions	18	7	6	22	8	16	10	10

Recognition results after adaptation:

Speaker	Word Accurcy	% Correct
1	71.14%	83.22%
2	69.33%	77.33%
3	84.47%	88.82%
4	76.03%	82.88%
5	80.93%	90.72%
6	82.98%	90.78%
7	83.33%	84.38%
8	84.00%	87.33%
Mean	79.03%	86.30%

Speaker	1	2	3	4	5	6	7	8
Number of words	149	150	161	146	194	141	96	150
Number of correctly recognized words	124	116	143	121	176	128	81	131
Deletions	2	5	1	7	1	2	9	7
Substitutions	23	29	17	18	17	11	6	12
Insertions	18	12	7	10	19	11	1	5

D.2.2 English

Recognition results before adaptation:

Speaker	Word Accurcy	% Correct
1	58.79%	69.78%
2	37.35%	40.96%
3	38.65%	42.33%
4	62.03%	70.89%
5	57.75%	66.84%
6	45.58%	55.10%
7	72.22%	83.33%
8	50.94%	56.60%
Mean	52.91%	60.71%

Speaker	1	2	3	4	5	6	7	8
Number of words	182	166	163	158	187	147	108	159
Number of correctly recognized words	127	68	69	112	125	81	90	90
Deletions	11	59	60	17	33	19	1	20
Substitutions	44	39	34	29	29	47	17	49
Insertions	20	6	6	14	17	14	12	9

Recognition results after adaptation:

Speaker	Word Accurcy	% Correct
1	81.87%	87.91%
2	75.30%	79.52%
3	74.85%	82.21%
4	77.85%	87.34%
5	82.89%	91.98%
6	80.27%	92.52%
7	87.04%	90.74%
8	85.53%	89.31%
Mean	80.70%	87.70%

Speaker	1	2	3	4	5	6	7	8
Number of words	182	166	163	158	187	147	108	159
Number of correctly recognized words	160	132	134	138	172	136	98	142
Deletions	6	11	8	1	7	1	0	2
Substitutions	16	23	21	19	8	10	10	15
Insertions	11	7	12	15	17	18	4	6

D.3 1G

D.3.1 Swedish

Recognition results before adaptation:

Speaker	**Word Accurcy**	**% Correct**
1	36.54%	46.15%
2	35.03%	42.37%
3	31.49%	34.25%
4	42.58%	59.35%
5	47.69%	54.36%
6	29.85%	48.51%
7	39.42%	50.96%
8	31.88%	35.00%
Mean	36.33%	46.37%

Speaker	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Number of words	156	177	181	155	195	134	104	160
Number of correctly recognized words	72	75	62	92	106	65	53	56
Deletions	25	35	52	12	24	8	14	34
Substitutions	59	67	67	51	65	61	37	70
Insertions	15	13	5	26	13	25	12	5

Recognition results after adaptation:

Speaker	**Word Accurcy**	**% Correct**
1	75.00%	78.21%
2	67.80%	77.40%
3	64.64%	70.17%
4	68.39%	77.42%
5	82.05%	90.77%
6	79.10%	88.81%
7	76.92%	80.77%
8	71.25%	72.50%
Mean	73.14%	79.50%

Speaker	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Number of words	156	177	181	155	195	134	104	160
Number of correctly recognized words	122	137	127	120	177	119	84	116
Deletions	12	6	20	8	3	2	7	17
Substitutions	22	34	34	27	15	13	13	27
Insertions	5	17	10	14	17	13	4	2

D.3.2 English

Recognition results before adaptation:

Speaker	Word Accurcy	% Correct
1	42.29%	42.29%
2	23.46%	23.46%
3	19.23%	19.78%
4	55.25%	55.25%
5	35.42%	37.50%
6	45.03%	47.68%
7	50.49%	50.49%
8	28.90%	30.06%
Mean	37.50%	38.31%

Speaker	1	2	3	4	5	6	7	8
Number of words	175	162	182	181	192	151	103	173
Number of correctly recognized words	74	38	36	100	72	72	52	52
Deletions	70	81	95	52	87	31	33	61
Substitutions	31	43	51	29	33	48	18	60
Insertions	0	0	1	0	4	4	0	2

Recognition results after adaptation:

Speaker	Word Accurcy	% Correct
1	67.43%	67.43%
2	65.43%	66.67%
3	56.59%	57.14%
4	72.93%	74.03%
5	60.42%	61.46%
6	80.79%	81.46%
7	72.82%	72.82%
8	54.91%	57.23%
Mean	66.30%	67.28%

Speaker	1	2	3	4	5	6	7	8
Number of words	175	162	182	181	192	151	103	173
Number of correctly recognized words	118	108	104	134	118	123	75	99
Deletions	45	33	54	38	61	13	21	31
Substitutions	12	21	24	9	13	15	7	43
Insertions	0	2	1	2	2	1	0	4

D.4 4G

D.4.1 Swedish

Recognition results before adaptation:

Speaker	Word Accurcy	% Correct
1	45.67%	50.96%
2	35.80%	41.98%
3	30.87%	32.21%
4	51.63%	61.96%
5	47.06%	53.85%
6	39.86%	47.97%
7	35.42%	41.67%
8	34.95%	43.55%
Mean	40.16%	46.77%

Speaker	1	2	3	4	5	6	7	8
Number of words	208	162	149	184	221	148	96	186
Number of correctly recognized words	106	68	48	114	119	71	40	81
Deletions	45	31	47	19	38	23	17	35
Substitutions	57	63	54	51	64	54	39	70
Insertions	11	10	2	19	15	12	6	16

Recognition results after adaptation:

Speaker	Word Accurcy	% Correct
1	72.12%	81.25%
2	59.88%	69.14%
3	67.11%	77.18%
4	76.09%	82.61%
5	79.19%	92.76%
6	78.38%	83.78%
7	75.00%	79.17%
8	66.67%	74.19%
Mean	71.80%	80.01%

Speaker	1	2	3	4	5	6	7	8
Number of words	208	162	149	184	221	148	96	186
Number of correctly recognized words	169	112	115	152	205	124	76	138
Deletions	13	13	9	10	3	4	7	11
Substitutions	26	37	25	22	13	20	13	37
Insertions	19	15	15	12	30	8	4	14

D.4.2 English

Recognition results before adaptation:

Speaker	**Word Accurcy**	**% Correct**
1	41.80%	41.80%
2	11.11%	12.70%
3	19.19%	20.93%
4	42.08%	42.62%
5	28.17%	28.17%
6	38.69%	39.29%
7	40.91%	41.82%
8	22.40%	23.50%
Mean	30.54%	31.35%

Speaker	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Number of words	189	189	172	183	213	168	110	183
Number of correctly recognized words	79	24	36	78	60	66	46	43
Deletions	84	123	76	71	110	54	44	64
Substitutions	26	42	60	34	43	48	20	76
Insertions	0	3	3	1	0	1	1	2

Recognition results after adaptation:

Speaker	**Word Accurcy**	**% Correct**
1	61.38%	62.43%
2	49.74%	49.74%
3	61.63%	62.79%
4	64.48%	65.03%
5	52.58%	53.99%
6	66.07%	67.86%
7	64.55%	64.55%
8	48.63%	50.27%
Mean	58.63%	59.59%

Speaker	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Number of words	189	189	172	183	213	168	110	183
Number of correctly recognized words	118	94	108	119	115	114	71	92
Deletions	56	58	38	55	79	31	27	45
Substitutions	15	37	26	9	19	23	12	46
Insertions	2	0	2	1	3	3	0	3

D.5 Higher G-loads

Recognition results before adaptation:

	Swedish	English
Number of words	155	152
Number of correctly recognized words	72	84
Deletions	24	22
Substitutions	59	46
Insertions	13	6

	Word Accurcy	% Correct
Swedish	38.06%	46.45%
English	51.32%	55.26%

Recognition results after adaptation:

	Swedish	English
Number of words	155	152
Number of correctly recognized words	103	101
Deletions	12	15
Substitutions	40	36
Insertions	25	12

	Word Accurcy	% Correct
Swedish	50.32%	66.45%
English	58.55%	66.45%

E List of Swedish phrases for the recordings in the Gripen aircraft

1	ledningsmod STRIL
2	ledningsmod radar
3	ledningsmod flygplan
4	fråga bränsle
5	destination brytpunkt tre
6	destination bas noll
7	öka kartskala
8	fråga hembränsle
9	minska sökavstånd
10	destination mål ett
11	markör MI
12	markör brytpunkt fyra
13	markör TI
14	sökvolym två
15	radio ett baskanal Adam
16	markör Filip Ivar
17	sambandsalternativ sju verkställ
18	radio två baskanal noll fyra noll Bertil
19	sökvolym ett
20	stega
21	sambandsalternativ två nio verkställ
22	radio ett frekvens ett fem åtta tre fem verkställ
23	markör Sigurd Ivar
24	radio två AM frekvens ett fyra sju fyra två fem avbryt
25	radio ett frekvens tre sex noll sex sju fem verkställ
26	radio två Caesar
27	radio ett baskanal två nio sex Caesar två
28	minska kartskala öka minska
29	öka sökavstånd minska minska
30	radio ett frekvens ett sex åtta ångra sju fyra fem verkställ

F List of English phrases for the recordings in the Gripen aircraft

1	guidance ground control
2	guidance radar
3	guidance air-command
4	report fuel
5	destination waypoint three
6	destination base zero
7	increase map scale
8	report bingo fuel
9	reduce radar range
10	destination target one
11	cursor right
12	cursor waypoint four
13	cursor centre
14	search volume two
15	radio one channel Alpha
16	cursor left
17	communication preset seven go
18	radio two channel zero four zero Bravo
19	search volume one
20	priority step
21	communication preset two nine go
22	radio one frequency one five eight three five go
23	cursor HUD
24	radio two AM frequency one four seven four two five abort
25	radio one frequency three six zero six seven five go
26	radio two channel Charlie
27	radio one channel two nine siz Charlie two
28	reduce map scale increase reduce
29	increase radar range reduce reduce
30	radio one frequency one six eight clear seven four five go