Database Adaptation for Speech Recognition in Cross-Environmental Conditions

Oren Gedge[1], Christophe Couvreur[2], Klaus Linhard[3], Shaunie Shammass[1], Ami Moyal[1]

[1]NSC – Natural Speech Communication
33 Lazarov St., Rishon-Lezion, Israel
{oreng, shaunie, ami}@nsc.co.il
[2]Scansoft
Guldenspoorenpark 2F, B-9820 Merelbeke, Belgium
christophe.couvreur@scansoft.com
[3]DaimlerChrysler AG
P.O.Box 2360, D-89013, Ulm, Germany
klaus.linhard@daimlerchrysler.com

Abstract

This study aims to simulate conditions that reflect the needs of speech-controlled consumer devices. In particular, it must be ascertained whether training in one type of environmental condition can be effectively adapted to other acoustic conditions, without having to perform costly collection in each specific type of environment. The adaptation tool performs two tasks: convolution of the clean speech signal with a given (room) Impulse Response (IR) and addition of noise to the convolved speech signal. Noise addition is done using recordings of typical environmental noise sources. Baseline, cross-tests and adaptation tests were performed. Results of the convolution and noise addition tests are presented for a speaker-dependent name recognition task. It is shown that adaptation reduces the recognition error rates when compared to the cross-tests. Ongoing tests within the SPEECON project are currently underway for evaluating the effectiveness of straight noise addition after convolution. For the speaker-independent case, preliminary tests on a database specifically collected for testing purposes have been performed.

1. Introduction

SPEECON is a project for creating spoken databases for 20 languages that includes a research component. The project is developed by an industrial consortium for the purpose of training speech recognition systems and promotes voice-controlled consumer applications such as control of television sets, video recorders, audio equipment, toys, information kiosks, mobile phones, palmtop computers and car navigation kits. As part of SPEECON's research program, this study aims to simulate conditions that reflect the needs of speech-controlled consumer devices.

It is well known that training in one environment and testing in another has the effect of decreasing speech recognition performance. At the same time, database collection is a costly endeavor. Thus, it is important to study whether adaptation techniques can be developed that would effectively reduce the number of database collections needed for various types of noise and acoustic environments, while maintaining reasonable speech recognition rates.

This study represents an initial phase in the SPEECON research program, indicating the potential of using adaptation algorithms on databases in various environmental conditions. Three experiments were performed. First, a speaker-dependent recognition experiment tested the effects of microphone type and distance from speaker. Second, adaptation to different microphones and distances was tested with and without noise addition. Third, a baseline speaker-independent recognition test was performed that tested the adaptation algorithm on isolated and connected digit recognition tasks as well as on a command and control task.

The following section examines the goals of the paper. In Section 3, the adaptation tool is described. The SDR experiments and results are presented in Section 4. SIR baseline tests are presented in Section 5. Overall discussion of the results and the direction for future research is given in Section 6.

2. Goal of the paper

The main objective of this study is to show the potential of using database transformation methods for adapting acoustic data to different environments. This is particularly crucial for real-life applications involving speech-controlled consumer devices. It needs to be seen whether a system trained in one environment can be adapted to other acoustic conditions without collecting speech data in each separate environment, a costly and laborious task. In particular, it would be advantageous to capitalize on close-talk recordings to enhance performance of ASR for target far-talk applications.

Thus, it is an important objective to develop an effective adaptation tool to be used for maintaining cost-effectiveness. The overall goal is to show whether such methods are effective in typical consumer applications and environmental conditions.

3. Adaptation Algorithm

3.1. General Description

The goal of the adaptation tool is to transform a database collected in a quiet environment (no noise) with a close talk microphone (no reverberation) into a noisy and reverberated "far-field" environment. Under the assumption that noise is additive and that the effect of room acoustics and microphone can be represented

by linear convolution, a database adaptation tool has been developed.

The effective adaptation tool performs two tasks: convolution of the clean speech signal with a given (room) IR and addition of noise to the clean speech signal (see Figure 1). The impulse response used for linear convolution is estimated from measurements made in real rooms using optimal IR identification techniques. Alternately, the tool also offers the possibility of generating synthetic impulse responses. The synthetic impulse responses are generated by a newly designed algorithm. The idea behind this alternative approach is to synthesize impulse responses that match a high-level description of the acoustic properties of a specific room impulse response such as reverberation time, early-to-late ratio, and global frequency characteristics. These properties can be either computed from real impulse responses (via the tool), or measured directly in a room using standard acoustical measurement equipment (e.g. a sound-level meter) and then used in the tool.

Noise addition is performed with recordings of typical environmental noise sources. Such recordings are performed as part of the SPEECON project. Once a noise file is available, it is scaled and added to the clean speech signal (after convolution thereof with an impulse response, if necessary) to get the desired SNR. The SNR or, alternately, the speech and noise level can be measured using standard procedures.

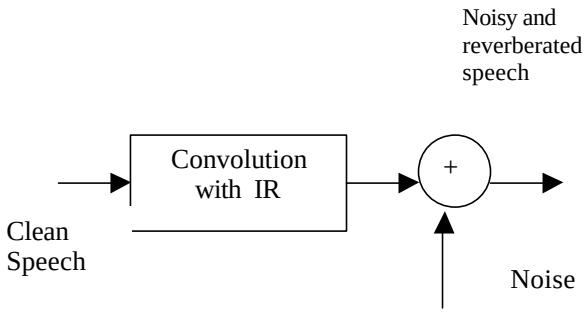

Figure 1: Operation of the adaptation tool

3.1.1. Motivation for the "Synthetic IR" approach

Using the optimal identification technique to identify the room impulse response should permit recovery as close as possible to the initial recognition rate with minimized mismatch between training and testing conditions, under the hypothesis that the IR used to adapt the training database matches exactly the IR used in the test environment.

In some circumstances, this requirement can lead to problems. First, with this approach, a new impulse response measurement is required for each new recording configuration (room, speaker and ASR system position). Since training a robust ASR system may require many IR's to cover the full spectrum of possible room configurations that can be encountered, the resulting data collection effort can rapidly become overwhelming.

Second, the optimally identified IR may be "too specific". That is, due to very precise identification, it will model a specific room and microphone/speaker configuration. Very small deviations from this configuration (e.g. moving the mouth of the speaker by a few centimers!) will result in changes of the IR, and therefore in a performance loss for the adapted ASR system, as has been observed by Couvreur, *et al* (2000). Some form of "smoothing" of the data is needed.

A possible solution to the first problem is to use an acoustic simulation package such as that in Rindel (2000). However, very high quality impulse response generators are often complex, expensive and match "too specific" configurations. Because they require minute descriptions of the geometry and acoustical properties of the room, they are not much cheaper than real measurements.

The approach we propose, namely to synthesize "random" IR's that match high-level properties of the room under consideration, can solve both problems. Since only high-level characteristics are taken into account, it is very easy to generate many IR's from inexpensive measurements. Furthermore, the fact that only high-level characteristics are matched provides a natural form of "smoothing". Of course, this requires that the high-level characteristics that are used are representative of the room as far as the operation of the ASR system is concerned.

3.2. Impulse Response Estimation, Analysis and Synthesis

3.2.1. General Principle

There are two main methods for obtaining an IR that can be used to convolve the clean speech signal (Figure 1). The first method is to use real measurements from microphones placed in various positions in the room. The second method is to generate synthetic IR's from parameters that capture high-level characteristics of the room. The latter parameters can be obtained in two ways: 1) from a real identified IR or 2) from a geometric and acoustic description of the room via a mapping.

In the following sections, these methods are outlined.

3.2.2. IR Estimation using Real Measurements

IR estimation is a well-known classical problem of room acoustics (Gardner, 1998), but for adaptation of speech databases the estimation problem needs some special considerations.

Typically in room acoustics for measuring the room IR one would measure between an electrical reference noise signal before sending this signal to a loudspeaker and the room microphone placed somewhere in the room. In this case, direct measurements between the close talk microphone of the speaker and the room microphone were taken in order to calculate the IR. This way we get the best acoustical match that also includes a small feedback from the room into the reference microphone.

The IR of a room is estimated by inputting a pink noise sequence recorded with both close-talk and far-talk microphones. IR identification is done using modules that include: 1) removing recording artifacts that would corrupt the IR estimation, 2) solving normal

equations for making the IR estimation optimal, 3) LMS filtering for offering tracking capabilities.

Figure 2 shows the block diagram for the IR identification software.

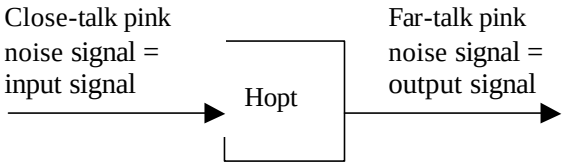

Figure 2: Optimal identification

Input correlation matrix:	Rxx
Cross-correlation matrix:	Rxy
Identified Impulse Response	Hopt

$Hopt = Rxz\ Rxx^{-1}$

Figure 3 shows the block diagrams for the N-LMS (normalized least mean squared) algorithm.

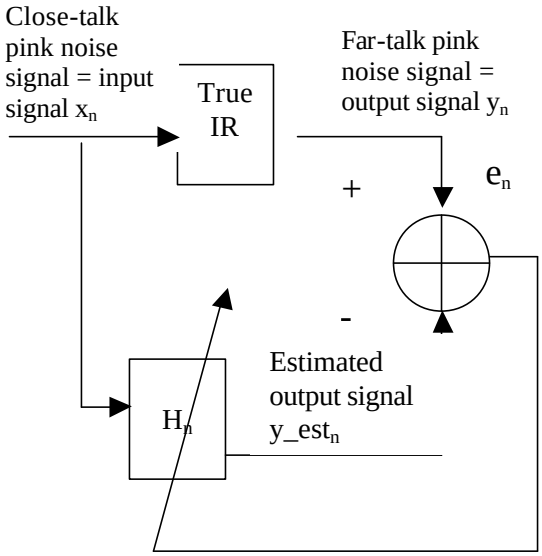

Figure 3: N-LMS identification

In this algorithm, the identification is done step-by-step. For every new sample, the transfer function is computed that converges to the optimal solution. To accelerate the convergence of the algorithm, it can be initialized with the optimal solution previously calculated.

The N-LMS algorithm is described below:

Input signal	x_n
Output signal	y_n
Estimated output signal	y_est_n
Error signal	e_n
Impulse Response order	N
Input signal vector	$X_n = (x_n, ..., x_{n-N+1})$
Identified impulse responses vector	$H_n = (h_{n,0}, \ldots, h_{n,N-1})$
Adaptation step-size	?
Normalized step-size	?_norm
Inverse input correlation matrix	P_n

Equations:

$y\_est_n = H_{n-1} * X_n$
$e_n = y_n - y\_est_n$
$P_n = P_{n-1} + x_n^2 - x_{n-N}^2$
?_norm = ? / P_n
$H_n = H_{n-1} + ?\_norm ? e_n ? X_n$
* = convolution operator

3.2.3. IR Synthesis Method using Modeling Software

The synthesis method used in this work is a variant of the one introduced by Couvreur & Couvreur (2000) and Couvreur, *et al* (2000). The idea is to generate impulse responses by manipulating (weighting and filtering) a white noise sequence. The manner in which the IR's are generated is summarized in Figure 4. The leftmost part of the figure summarizes the different steps in the algorithm. The rightmost column gives an example of IR generated by this process at the different steps.

The synthesis process starts with a white noise sequence (pseudo-random). A non-linear downsampling operation is then performed to ensure that the proper density of reflections is present in the room IR (Gardner, 1998). The downsampled sequence is then filtered by a LPC filter that represents main resonant modes of the room. Note that the resonance modes are characteristic of the room, not of a specific position in the room. As an alternative, the software also allows the frequency spectrum of the original IR to be used instead of an LPC model (h2). An exponentially decaying envelope is then used to modulate the amplitude of the noise sequence (h3). The decay time of this envelope is directly linked to the reverberation time of the room. Some gain and early-late energy normalizations are performed to adjust for different reverberation vs. direct path situations (h4). Finally, the IR is high-pass filtered to remove DC artifacts (h5).

The IR synthesis software is driven by a set of parameters (for the LPC coefficients, the exponential decay, energy and gain normalizations). All the responses generated by this software are computed using only parameters stored in a file, which is the output of the analysis/extraction parameters software. The analysis software takes as input an IR identified by the identification software described in Section 3.2.2.

This software outputs six IR models. Within these models, one subset of 3 responses uses only the noise sequence generator and the other subset applies the sparseness filter. This allows the user to experiment with the effects of the various modeling components. The following models are thus produced by the software:

Model 0: white noise sequence generator + envelope,

Model 1: white noise sequence generator + LPC information + envelope,
Model 2: white noise sequence generator + filtering with the real IR + envelope,
Model 3: white noise sequence generator + sparseness + envelope,
Model 4: white noise sequence generator + sparseness + LPC information + envelope,
Model 5: white noise sequence generator + sparseness + real IR + envelope.

The block diagram is shown in Figure 4.

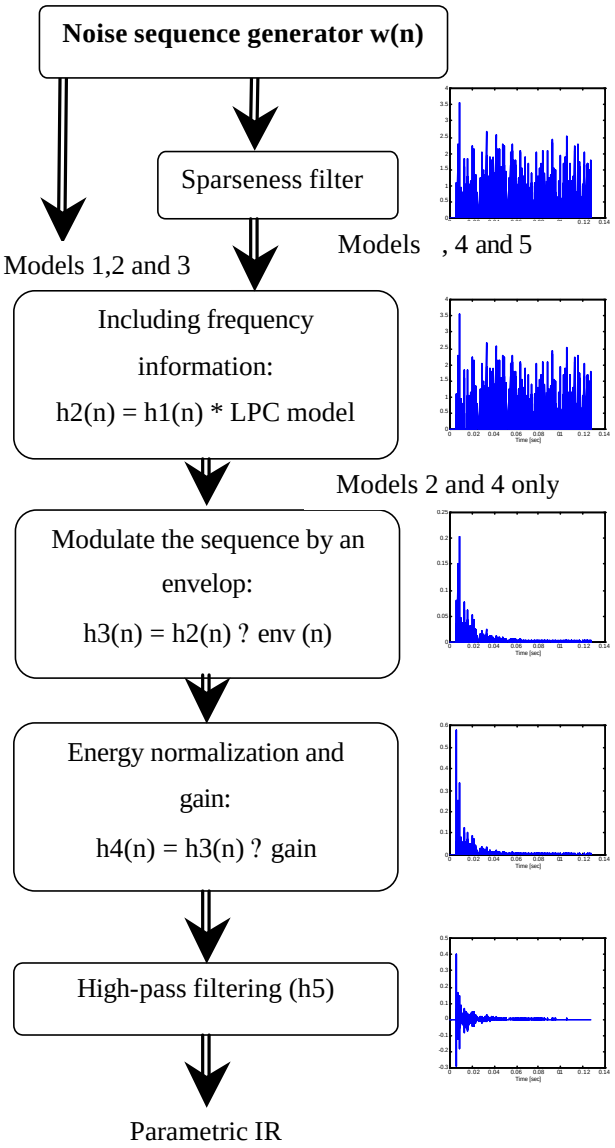

Figure 4: Global Block Diagram

3.3. Convolution

Convolved/reverberated speech is obtained by convolving speech recorded at the close-talk microphone with the IR's calculated for the other microphone positions.

In the tool, this convolution is performed efficiently in the frequency domain using an OverLap-Add (OLA) method (Oppenhiem and Schaffer, 1999). This approach is preferred over the plain linear convolution because of the length of typical room impulse responses.

3.4. Noise Addition

Noise addition is done by adding a recorded noise sequence to the clean speech utterance. The recorded noise sequence must be representative of the target operating environment for the system. Such noise recordings are part of the SPEECON data collection effort.

The noise addition algorithm is as follows:

x(k)	initial clean speech utterance
n(k)	interference noise recording
y(k) = x(k) + g ? n(k)	noisy speech utterance

The noise gain, g, is chosen in order to obtain the desired SNR statistics.

Special attention is paid to the addition of the noise recording. The tool is intended for use with very long noise recordings (up to 30 minutes) when compared to the typical clean speech utterance. When batch-processing a large series of speech utterances, the noise addition tool updates a pointer to ensure that the full noise recording is used and not always the same small initial segment.

4. Speaker-Dependent Experiments

4.1. General Overview

This section describes database development for the purpose of testing the potential for the adaptation procedure. The recognition was performed using NSC's speech recognition engine, NSCEngine, removing all tools for robustness. In the first section, the databases themselves are described. The following sections describe speaker dependent tests done on these databases.

4.2. Speaker Dependent Recognition Experiment

4.2.1. Method

As a pilot case, two Hebrew speakers were recorded with two microphone types (Cardioid and Omni) in several positions (close/middle/far). The speakers were required to read 40 Hebrew application words with ten repetitions. A loudspeaker playing pink noise was placed near the close cardioid microphone at the same position as the speaker. Pink noise recorded from the speaker's mouth level was used for calculating the impulse response (IR).

Three types of tests were done: baseline tests, cross tests and convolution tests, as described below.

Baseline tests are done using a test set of speech files that were recorded with the same microphones used in training. Thus, in the baseline tests, training and testing are done in the same environment.

Cross-tests are done by training the speech recognizer with speech data recorded from the close cardioid microphone and testing with speech recorded

using the other microphones in the test setup. This represents the worst case scenario, where the difference between training and testing conditions are maximal.

The effectiveness of adaptation is tested in *convolution tests*, which involve training with convolved speech while testing on files from another environment. In this case, the speech recorded with the close cardioid microphone is convolved with the IR calculated for the other distances, while testing is done using the original files. The difference between results obtained by the cross tests and the convolution tests represents the effectiveness of the adaptation tool.

4.2.2. Results

Results are shown in Table 1. As can be seen in the table, convolution improved the recognition results, particularly for the far microphone case.

Baseline Tests		Cross Tests		Convolution Tests	
Close	95.9%	---	---	---	---
Mid	92.5%	Mid	96.6%	Mid	97.5%
Far	91.8%	Far	94.1%	Far	96.6%

Table 1: Results for Speaker Dependent Experiment

Further investigation shows that the higher recognition rates in the cross tests were due to the voice activity detection (VAD) function.

4.3. Noise Addition Experiment

4.3.1. Method

The recording setup is shown in Figure 5.

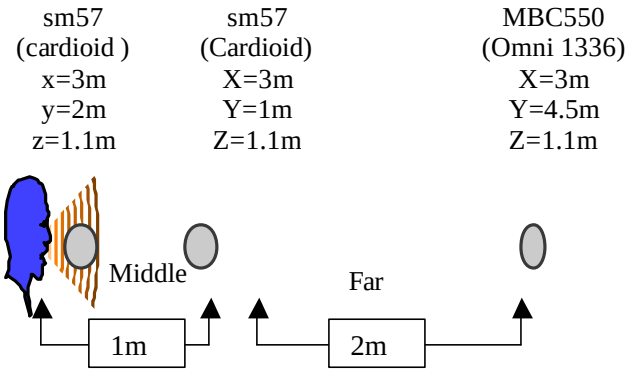

Figure 5: Recording Setup

Two native English speakers were recorded, one female and one male, using four microphones. These speech samples were adapted to the middle and far microphones and were then tested in various noisy environments. Noise was added to the convolved files, and the speech samples were then tested in the same noise environment that had been added.

Two types of environmental noises were added: computer room noise and noise from a shopping mall. As in the other experiments, baseline, cross-tests and convolution tests were done.

4.3.2. Results

Results are shown in Table 2. Results show that convolution generally improves recognition, especially in the highly noisy environment and using the far microphone. Word error rates (WER's) decreased between 27%-40% when convolution was performed.

Convolution with noise addition does not improve recognition rate, while noise addition to the clean speech from the close cardioid microphone improves recognition rate.

Mic's	Baseline Tests	Cross Tests	Conv Tests	Conv & Noise addition
Middle	90.84%	84.93%	87.20%	85.17%
Far	78.92%	59.32%	69.01%	67.32%

Table 2: Results for Noise Addition Experiment

Figure 6 shows the results of the noise addition in different target SNR's. As can be seen, noise addition to the clean speech from the close cardioid microphone improves recognition rate.

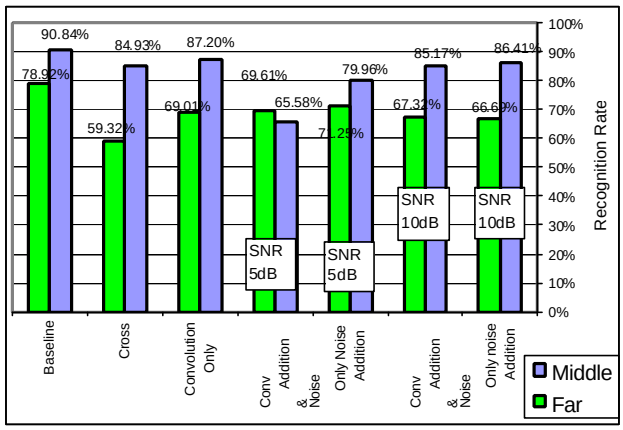

Figure 6: Results for Noise Addition Experiment

5. Speaker Independent Experiments

5.1. General Overview

This section describes database development for the purpose of testing the potential for the adaptation procedure in the speaker-independent (SI) case. The recognition was performed using a simplified version of one of ScanSoft's recognition engines. In the first section, the SI database is described. The following sections describe the experiments conducted and present some preliminary results.

5.2. Validation SI Database

The Validation SI Database was collected for the purpose of evaluating the adaptation tool described in Section 3 with a speaker-independent recognizer.

A French database was collected in the area of Paris (Ile de France). A total of 46 speakers were recorded in 137 sessions in different recording conditions. Two rooms were used: a small room (office) and a large room (meeting room). Noise conditions were varied by opening or closing a window on a busy city street, considered as 'noisy' and 'quiet' environments, respectively. In some of the recording sessions, the speaker was moving ('dynamic') in order to show the impact of dynamic variation of the acoustic path between the speaker and the microphones (IR).

The recording set-up is similar to the one shown in Figure 5. It uses three microphones: a close-talk/headset microphone, a far-talk cardioid microphone at a medium distance, and another far-talk omnidirectional microphone at the opposite end of the room. Figure 7 illustrates the A-weighted segmental biased SNR in the various room and recording conditions.

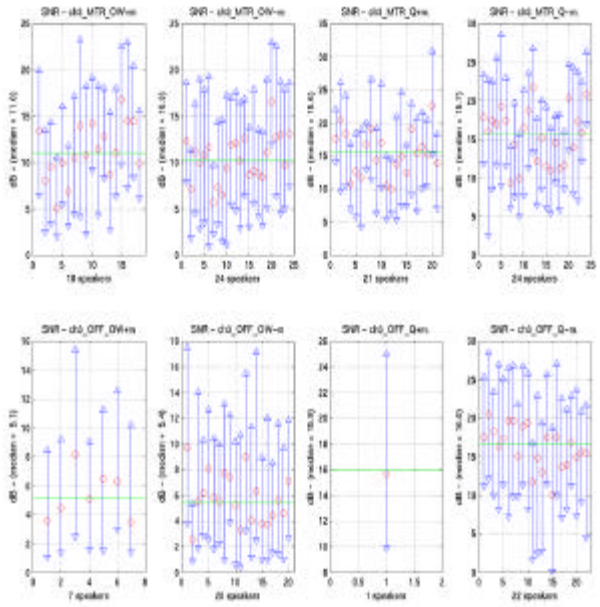

Figure 7: A-weighted SNR in Various Room and Recording Conditions

In each of the recording sessions, the speaker was asked to utter 10 isolated command & control words (out of a set of 141), 10 isolated digits, 10 free-form sequences of 10 connected digits, and 5 sequences of 4 free-form digits. For IR identification, pink noise was also played by a loudspeaker at the speaker's head location and recorded. Noise background recordings were also performed. The recording sessions also included 20 minutes of background noise.

5.3. ASR System and Training Path

The experiments are conducted with a SI phonetic recognizer trained on a French corpus of about 80 hours of clean speech (close talk microphone recorded in office conditions). The recognizer is a simplified version of one of ScanSoft's embedded recognition engines.

The "robustification" of the engine is performed by training it on the same 80-hour corpora processed by the simulation tool described in Section 3. This is done for the various reverberation and noise conditions covered by the Validation SI database described in the previous section. The training path makes use of synthetic impulse responses starting with an identified IR from the pink noise recordings. The synthetic IR's are randomized. That is, multiple synthetic IR's are generated for the target conditions, and randomly shuffled during the database convolution (see Couvreur, *et al*, 2000). For noise addition, the level of the speech and noise signal is adjusted to align the mean SNR on the processed database and the SNR measured on the Validation SI database. The training path also allows for the combination of multiple reverberation and noise conditions in the training (multi-style training) in the hope of getting a recognizer that will be able to operate in different environments.

5.4. Experimental Results

Recognition experiments are performed on the Validation SI database. An isolated digit grammar with the 10 French digits is used with the "digit" part of the database. A free-length connected digit grammar is used with the connected-digit part of the database. A command and control grammar with 141 commands (the 10 recorded ones "enriched" to reach 141) is used with the command and control part of the database.

Baseline results are given in Tables 3, 4 and 5 for isolated digit, connected digit and C&C recognition tasks, respectively. (Note: In the following tables, MIR stands for "meeting room", OFF for "office", Q for "quiet room", and OW for "open window".)

Rec Condition	SER(%)	WER(%)	#UTT	#WRD	#SPKRS
	Channel 1				
MIR-OW	0.2	0.2	450	450	42
MIR-Q	0.7	0.7	510	510	45
OFF-OW	0.4	0.4	400	400	27
OFF-Q	0.7	0.7	390	390	23
	Channel 2				
MIR-OW	1.2	1.2	450	450	42
MIR-Q	0.2	0.2	510	510	45
OFF-OW	1.1	1.1	400	400	27
OFF-Q	2.0	2.0	390	390	23
	Channel 3				
MIR-OW	24.2	24.2	450	450	42
MIR-Q	7.6	7.6	510	510	45
OFF-OW	45.4	45.4	400	400	27
OFF-Q	9.6	9.6	390	390	23

Table 3: SI Baseline Recognition Rates, Isolated Digits

Rec Condition	SER(%)	WER(%)	#UTT	#WRD	#SPKRS
	Channel 1				
MIR-OW	2.3	1.9	450	540	42
MIR-Q	2.4	2.0	510	612	45
OFF-OW	5.2	4.6	400	480	27
OFF-Q	4.3	3.6	390	468	23
	Channel 2				
MIR-OW	5.5	4.6	450	540	42
MIR-Q	3.7	3.1	510	612	45
OFF-OW	6.1	5.4	400	480	27
OFF-Q	6.1	5.1	390	468	23
	Channel 3				
MIR-OW	38.8	36.1	450	540	42
MIR-Q	33.8	30.7	510	612	45
OFF-OW	68.3	69.8	400	480	27
OFF-Q	22.4	19.6	390	468	23

Table 4: SI Baseline Recognition Rates, Connected Digits

Rec Condition	SER(%)	WER(%)	#UTT	#WRD	#SPKRS
	Channel 1				
MIR-OW	12.7	1.7	90	720	5
MIR-Q	10.0	1.7	90	720	4
OFF-OW	12.4	1.5	330	2640	17
OFF-Q	7.7	1.1	315	2520	13
	Channel 2				
MIR-OW	16.0	2.2	90	720	5
MIR-Q	6.7	0.8	90	720	4
OFF-OW	21.4	3.1	330	2640	17
OFF-Q	11.5	1.6	315	2520	13
	Channel 3				
MIR-OW	95.3	52.6	90	720	5
MIR-Q	56.7	11.4	90	720	4
OFF-OW	97.8	65.0	330	2640	17
OFF-Q	56.1	11.8	315	2520	13

Table 5: SI Baseline Recognition Rates, Command and Control Words

As can be seen, the reverberation and noise conditions for the far-talk microphone are particularly harsh, leading to very high error rates. The mid-distance microphone seems to be a more realistic target. Based on the results obtained in the SDR experiments of Section 4, these preliminary results suggest that a performance gain of 20 to 40% relative is possible for this mid channel by using convolution and noise addition.

6. Discussion

The potential for adaptation is shown in this initial phase of the SPEECON research. It is shown that adaptation methods involving convolution improves recognition performance. In particular, convolution is highly recommended for the far microphone and for highly noisy environments in the speaker dependent case.

Further testing of the potential of the adaptation toolbox needs to be done on the above speaker-independent database before performing evaluation on the larger SPEECON databases. Future work includes evaluating the adaptation methods on the large-scale databases collected in various acoustic environments within the SPEECON project using several hundreds of speakers.

7. References

Couvreur, L., & Couvreur, R. (2000). On the use of artificial reverberations for ASR in highly reverberant environments. Paper presented at the 2nd IEEE Benelux Signal Processing Symposium, Hilvarenbeek, The Netherlands, March 23-24, 2000.

Couvreur L., Couvreur C. & Ris, C. (2000). A Corpus-based Approach for Robust ASR in Reverberant Environments, In Proceedings of the 6th International Conference on Spoken Language Processing (Vol. 1: pp. 397-400) Beijing, China, Oct. 2000.

Rindel, J.H., (2000). The Use of Computer Modelling in Room Acoustics, Journal of Vibroengineering, Vol. 3, (4).

Gardner, W.G. (1998). Reverberation algorithms, In M. Kahrs and K. Brandenburg (Eds.), Applications of Digital Signal Processing to Audio and Acoustics, (pp. 85-132) NY: Kluwer.

Oppenheim, A.V., and Schaffer, R. W. (1999). *Discrete-Time Signal Processing*, 2d ed., NJ: Prentice-Hall.

Acknowledgements

This project is partly funded by the European Commission as part of the SPEECON project.