Database Adaptation for ASR in Cross-Environmental Conditions in the SPEECON Project

Christophe Couvreur[†], *Oren Gedge*[⋆], *Klaus Linhard*[‡]
Shaunie Shammass[⋆], *Johan Vantieghem*[†]

[†]ScanSoft Belgium, Guldensporenpark 32, B-9820 Merelbeke, Belgium
[⋆]NSC – Natural Speech Communication, 33 Lazarov St., Rishon-Lezion, Israel
[‡]DaimlerChrysler AG, P.O.Box 2360, D-89013, Ulm, Germany

Abstract

As part of the SPEECON corpora collection project, a software toolbox for transforming speech recordings made in a quiet environment with a close-talk microphone into far-talk noisy recordings has been developed. The toolbox allows speech recognizers to be trained for new acoustic environments without requiring an extensive data collection effort. This communication complements a previous article in which the adaptation toolbox was described in details and preliminary experimental results were presented. Detailed experimental results on a database specifically collected for testing purposes show the performance improvements that can be obtained with the database adaptation toolbox in various far-talk and noisy conditions. The Hebrew corpus collected for SPEECON is also used to assess how close a recognizer trained on simulated data can get to a recognizer trained on real far-talk noisy data.

1. Introduction

SPEECON [1], launched in February 2000, is a project focusing on collecting linguistic data for speech recognizer training for consumer electronics (CE) devices. SPEECON is funded as a shared-cost project by the European Commission's Information Societies Technologies (IST) Programme. The SPEECON industrial consortium is collecting speech databases for 26 languages or dialectal zones in typical environments for CE applications. As part of the SPEECON project, a software toolbox for transforming speech recordings made in a quiet environment with a close-talk microphone into far-talk noisy recordings has been developed. The goal of the toolbox is to allow speech recognizers to be trained for new acoustic environments without requiring an extensive data collection effort. The SPEECON toolbox algorithm and early speech recognition results obtained with it have been presented in [2]. In this communication, we present extensive experimental results obtained with the toolbox on two test databases: a dedicated French database collected in various environmental conditions with a series of microphones, and the Hebrew SPEECON database. The latter also allows us to assess how close a recognizer trained on simulated data can get to a recognizer trained on real far-talk noisy data for CE environments.

The paper is organized as follows. In section 2, we briefly describe the SPEECON toolbox and how it can be used in practice. Sections 3 and 4 describe the experiments conducted on the dedicated French database and on the Hebrew SPEECON database, respectively. Finally, we draw conclusions in section 5

2. The SPEECON Toolbox

2.1. The SPEECON Environment Recordings

The SPEECON data collection effort takes into account the acoustic variability that may be encountered in a CE scenario. That is, recordings are made with multiple microphone configurations, in different room types, in different locations within these rooms, and with various noise backgrounds. In addition, specific items are recorded for use with the SPEECON Toolbox: pink noise and maximum-length sequences (MLS) for room impulse response (IR) identification and noise background. More details on the SPEECON recording scenarios can be found in [4], publicly available on [1].

2.2. Description of the Toolbox

The SPEECON Toolbox consists of two parts. The first part, developed at DaimlerChrysler, supplies tools for the estimation of the acoustical characteristics of a room. In particular, this Room Acoustic Toolbox can:

1. Estimate room IR's from MLS-sequence recordings, from speech recordings or from pink noise recordings with various IR identification techniques (LMS, time-domain correlation, frequency-domain correlation);
2. Measure room acoustic properties (e.g. reverberation times).

The second part, developed at ScanSoft, supplies tools for the adaptation of clean speech databases to other acoustic conditions. The Adaptation Toolbox can:

1. Extract a set of acoustically meaningful parameters from a room IR;
2. Generate synthetic IR's that match the acoustic parameters of a room by a method similar to the one proposed in [3];
3. Convolve a clean speech signal with (artificial or measured) room IR's;
4. Add noise to clean speech recordings with scaling adjustments to obtain the desired SNR characteristics.

The two parts of the toolbox interface together and use as their input the special recordings made during the SPEECON data collection effort described in section 2.1. A detailed description of the SPEECON toolbox can be found in [2].

2.3. Using the Toolbox in Practice

It is well known that speech recognizers perform best when the training data matches the operating conditions (SNR, microphone position, etc). When recognizers have to be used in

a mix of environments, the common practice is to use mixed-style training with matched data for the various target environments. The Toolbox can be used to perform noise addition and IR convolution on "standard" close-talk clean speech recordings to map them to a target acoustic environment. The toolbox is designed to offer a lot of flexibility on the noise and reverberation simulation. However, in order to obtain maximum accuracy, the noise recordings used for noise addition and the IR's used for convolution must be carefully selected. Their integration in the training process of the speech recognizer is critical too. We found the following guidelines to work well in practice.

When performing the noise addition, the noise recording(s) and the scaling parameters should be selected to match the noise characteristics of the target environment: SNR and noise types. For the SNR, the speech and noise scaling factors should not be set individually per utterance so that every noisy utterances has the same target SNR. Instead, the scaling factor should be such that the noisy database presents the same SNR *statistics* (mean, variations) as the target environment.

When performing convolution, more than one room IR should be used. Room IR's are very sensitive to room geometry, speaker and microphone positions, temperature, etc. To avoid overtraining, multiple IR's should be used. We found that 10-15 IR's were sufficient to give good results for one room category. The range of IR's used should not only cover both the range of rooms of interest for the target environment, but also the variability inside a room (microphone/speaker positions and other influencing factors). When multiple IR measurements are made, as is the case in the SPEECON database, the IR's identified with the Room Acoustics toolbox can be used directly for convolution. To ensure proper variability in the training data, the convolution should be randomized. That is, different IR's should be used for different speakers/sessions/utterances depending on the level of variability that is desired. When only a limited number of IR measurements are available (e.g. because no extensive recordings could be made), additional synthetic IR's with similar acoustic properties can be generated via the analysis and synthesis tools of the Adaptation toolbox. These synthetic IR's can then be randomized as regular IR's.

3. Preliminary Experiments

A first series of experiments has been conducted by ScanSoft as part of the research work package on adaptation included in the SPEECON project.

3.1. Database Description

A special purpose test database has been recorded for the evaluation of the database adaptation method during the research phase of the SPEECON project. This database, referred in the sequel as the Validation database, has been recorded in various conditions according to the standard SPEECON procedure [4]. It contains a total 137 sessions recorded by 46 native French speakers. Each of the sessions contains 20' of noise background recording, 10 isolated digits, 10 sequences of 10 digits, 5 sequences of 4 digits, and 10 command words (such as *stop*, *suivant*, etc). Sessions have been recorded either in an office or in a meeting room both in quiet and noisy conditions. The noisy conditions were obtained by opening a window on a busy street. In addition, pink noise sequences have been recorded for room IR identification. Three channels have been recorded: close-talk microphone (CT), desktop-mounted cardioid microphone (medium-talk or MT), far-talk omnidirectional microphone (FT).

It must be noted that the far-talk recordings present particularly aggressive conditions for ASR: the speech is heavily reverberated and the segmental A-weighted SNR [5] range between 10 dB(A) and 5 dB(A). This explains the low baseline accuracy obtained in these conditions. The SNR of the MT recordings is between 20 dB(A) and 25 dB(A).

Table 1: *Influence of noise addition on word accuracy.*

Testing	Training	ISD	CDD	C& C
office FT mike noisy	CT	50.7	34.8	30.6
	Noisy 2	50.5	36.1	33.0
	Mix	57.3	40.0	37.2
	Noisy 1	58.7	45.3	38.0
	Car	51.7	38.3	38.7
meeting room FT mike noisy	CT	74.6	47.6	55.4
	Noisy 2	76.2	51.6	57.9
	Mix	77.3	50.4	56.2
	Noisy 1	78.7	55.5	56.2
office CT mike noisy	CT	99.0	98.9	96.7
	Noisy 2	98.7	98.9	97.6
	Mix	99.0	98.8	96.9
	Noisy 1	98.5	98.0	96.7

3.2. Experimental Set-up

For the experiment, a standard "embedded" recognizer has been used. The recognizer uses phonetic models (HMM's). The front-end is MFCC-based and works at 11 kHz, with a frame-shift of 10 ms. It outputs 12 cepstral coefficient and their first- and second-order derivatives. The HMM's are discrete HMM's of generalized triphones speech units. The recognizer is trained on standard French corpora of clean close-talk recordings (BDSONS, BREF, etc). The training material amounts to about 80 hours of speech. The Room Acoustic and the Adaptation Toolboxes have been used to adapt the original close talk recordings to the test conditions.

The performance is measured on 3 tasks: isolated digits, connected digits, and 141 command words (the set of 10 recorded words enriched with additional commands).

3.3. Results and Analysis

We first evaluated the effect of noise addition alone. Noise addition was performed on the training corpora with the noise recordings made during the collection of the Validation DB. The noise-scaling factor was adjusted so that the SNR statistics of the adapted noisy data would match the SNR statistics of the Validation DB in two noise conditions. The low SNR condition, referred to as "Noisy 1," corresponds to the far-talk microphone with open window; its target SNR for training is 12 dB(A). The high SNR condition, referred to as "Noisy 2," corresponds to the desktop microphone; its target SNR for training is 25 dB(A). Recognizers were trained on noisy data alone for the "Noisy 1" and "Noisy 2" conditions and on a mix of data (mixed-style training) with 50% of clean data, 25% of low SNR data, and 25% of high SNR data. In order to assess the importance of using noise of a type similar to the noise at recognition time, the recognizer has also been retrained with noisy data obtained by adding car noise to the clean speech database. The car noise recording was made inside a regular sedan car driving at high speed. This noise has different characteristics than the background noise present in the Validation SI DB recordings, which is the non-stationary noise of a busy street with cars and trucks passing by and people discussing.

Table 1 summarizes the results obtained for the different tasks for various training and testing configurations.

The following observations have been made:

- Noise addition improves the accuracy on the noisy FT

Table 2: *Influence of convolution with synthetic IR's on word accuracy.*

Testing	Training	ISD	CDD	C& C
office FT mike quiet	CT	86.8	85.4	82.5
	mixMT	88.6	90.4	88.4
	mixMTFT	88.9	90.0	86.8
	mixFT	88.6	88.9	86.6
meeting room FT mike quiet	CT	90.2	86.6	63.3
	mixMT	91.5	89.3	74.8
	mixMTFT	91.6	89.0	72.9
	mixFT	91.7	88.2	69.8
office CT mike quiet	CT	99.5	98.2	99.3
	mixMT	99.3	97.3	99.1
	mixMTFT	99.3	97.3	98.9
	mixFT	99.0	96.8	98.7

Table 3: *Influence of the combination of convolution with IR's and noise addition on word accuracy.*

Testing	Training	ISD	CDD	C& C
office FT mike noisy	CT	50.7	34.8	30.6
	Noise	53.7	40.0	37.2
	Conv	52.7	40.5	39.7
	Conv+Noise	54.7	43.6	42.0
meeting room FT mike noisy	CT	74.6	47.6	55.4
	Noise	77.3	50.4	56.2
	Conv	77.7	55.1	63.7
	Conv+Noise	79.2	59.4	62.7
office CT mike noisy	CT	99.0	98.9	96.7
	Noise	99.0	98.8	96.2
	Conv	99.0	98.6	96.2
	Conv+Noise	99.0	98.7	96.2

channel recordings, especially when noise conditions are matched during training and testing;

- Noise addition degrades performance when tested on high SNR signal (if the SNR is not matched);
- Mixed-style training yields the best trade-off between the two conditions;
- Using a different noise type is not as good as using the target noise type.

Next, we investigated the effect of convolution only. We only considered the "quiet recording" part of the database to focus on the modeling of room reverberation. Pink noise recordings made in the quiet conditions in the office room have been used for the IR identification. The identified IR length was 2048 taps at 11 kHz sampling frequency. The room IR's have been identified with the Room Acoustics toolbox. The identified IR's were then analyzed using the Adaptation toolbox for both far-talk and medium-talk conditions. Based on the output of the analysis, 15 synthetic IR's were generated in accordance with the procedure described in section 2.3 for each microphone location. This to avoid "over-adaptation" to a specific IR. The 2×15 synthetic IR's were then convolved with the close talk training data. Recognizers with various mix of reverberation conditions were trained with the resulting data: close-talk recordings alone (CT), i.e. the original data; mixed-style training with 50% CT, 50% MT microphone (mixMT); mixed-style training with 50% CT, 50% FT microphone (mixFT); mixed-style training with 50% CT, 25% MT and 25% FT (mixFTMT). The IR were randomized. That is, a different IR was randomly picked for each speaker in the training set.

Table 2 gives the results obtained with the FT microphone in quiet acoustic conditions for the office room and meeting room recordings, and the results obtained for the CT recordings.

The following observations have been made:

- Convolution with synthetic IR improves accuracy on reverberated speech recordings, especially for FT recordings;
- With the modeling/synthetic IR approach, results are not too sensitive to the exact room conditions (i.e. using IR corresponding to the office recordings also helps with meeting room recordings);
- Mixed-style training yields the best trade-off;
- Using very high-level of reverberation (corresponding to the far-field case) during training does not necessarily yield the best results on very reverberated test data. This can probably be explained by a "model blurring" problem: if the noise and/or reverberation level is too high, the modeling power of the acoustic model is spent on the noise/reverberation part of the signal instead of on the discriminative speech features. That is, the recognizer "learns" the reverberation, not the speech.

In the final series of experiments, we combined noise addition and convolution with synthetic IR. For these experiments, we merged the settings of the previous two series of experiments. That is, we trained models on data combining convolution with a randomized synthetic impulse response and noise addition for two target environments: (1) office, MT microphone, and noisy conditions (open window, 12 dB(A) SNR); (2) office, FT microphone, and noisy conditions (open window, 25 dB(A) SNR). In order to be able to compare to the results obtained with models trained with noise addition only or with convolution only, we trained recognizers with the following mixed-style training configurations: 50% CT recordings, 25% simulated MT recordings and 25% simulated FT recordings.

Table 2 gives the results obtained with the FT microphone in noisy acoustic conditions for the office room and meeting room recordings, and the results obtained for the CT recordings.

The following observations have been made:

- Both convolution and noise addition yield improvements on far-field noisy data;
- The improvements are cumulative (on far-field noisy recordings), but not additive;
- The approach is robust to cross-room effects;
- Of course, the price to pay for the improved performance of the models on far-field noisy data is a small loss of performance on close-talk and/or clean recordings.

4. Experiments on the SPEECON Hebrew database

The toolbox developed by DaimlerChrysler and ScanSoft has been independently validated by NSC on the Hebrew SPEECON database after completion of the recordings. Since the SPEECON database includes enough speech utterances to train a recognizer and the recordings are made in noisy and reverberated conditions, it is possible to train a recognizer with real noisy far-talk data. Its performance can then be compared with the performance of a recognizer trained on simulated data obtained by the toolbox from the close-talk data. More specifically, in this series of experiments, the office recordings of the Hebrew SPEECON database have been used to train and test a

speaker-independent digit recognizer. The following tests were performed:

1. *Baseline*: Train and test at the same environment using the same microphone, i.e. matched conditions;
2. *Cross Test*: Train at CT microphone and test at target microphone (MT or FT);
3. *Adaptation Test*: Train at adapted CT microphone and test at target microphone (MT or FT).

4.1. Database Description

The data was recorded in five offices. Recordings of 146 speakers from three offices were used for training, and the recordings of 54 (distinct) speakers from the two other offices were used as a test set. The utterances of each speaker were recorded simultaneously using four microphones that were located in four different positions: CT headset, CT lavalier, MT located at 0.5 meters from the speaker, and FT located 2-3 meters from the speaker. The speakers were located in one of up to five different positions in the room. For each position, pink noise was recorded in order to enable calculation of reverberation characteristics. In addition to putting the loudspeaker where the speaker's head would be, positions of approximately 50 cm to the right and 50 cm to the left were also taken. So, it is possible to identify 3×5 IR's per room.

For each speaker, a recording of about twenty seconds of an exemplary noise sequence before the actual recording was done for noise addition.

4.2. Experimental Set-up

All experiments were performed using the standard HMM HTK Toolkit (HTK v3.1) [6]. It is configured to use 39 MFCC coefficients (cepstrum + first- and second-order derivatives) with a 16 ms frame shift. Continuous density, 8 Gaussians per mixture models were trained for the ten Hebrew digits with 10-12 states per model. The baseline performance of the recognizer when trained and tested on the CT channel is 96.4% word accuracy.

The convolution with the IR's for the target environment has been performed with real identified IR's. Since there was a sufficient number of real identified IR's, it was not needed to generate additional synthetic IR's as in section 3. The IR's used for the convolution had a length of 2048 samples. The IR's were scaled to preserve the energy of the speech signal.

Noise addition was performed with the noise recordings made during the recording session of the training speaker on the proper microphone (MT or FT).

4.3. Results and Analysis

Figure 1 summarizes the average recognition results for the MT and FT microphones. The results show that:

- Convolution improves the recognition rate significantly for both MT and FT microphones;
- Noise addition on convolved speech usually further improves the recognition rate, especially for the FT microphone; recognition rates achieved this way are very close to the baseline recognition rate;
- Noise addition on clean speech does not improve the recognition rate compared to results obtained in the cross tests.

This last result is in contradiction with the findings on the Validation database. The reason for the different observation might be that the noise level is much lower in the Hebrew SPEECON recordings.

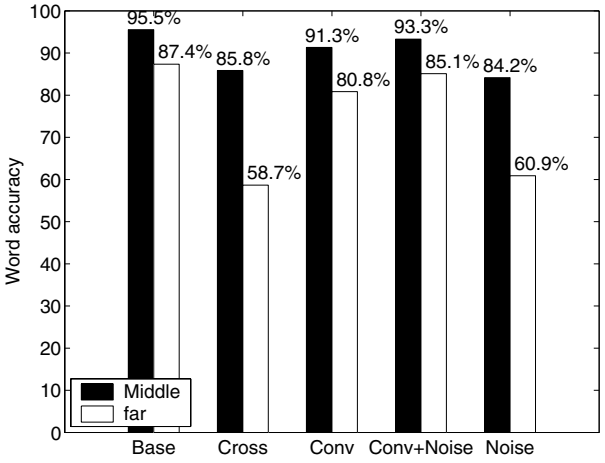

Figure 1: *Average results for the middle- and far-talk microphones for the Hebrew SPEECON Database.*

5. Summary

Summarizing the results obtained on the two databases, we can conclude that the simulation technique proposed here is effective, even if it does not allow a recognizer to match the performance of a recognizer trained on real data. In partice, we observed that noise addition helps improving performance when the SNR is bad, with the best improvements when the SNR and the noise type match the test situation, and that convolution helps improving performance for far-field recordings (MT or FT). The accuracy gains of noise addition and convolution are cumulative when dealing with far-talk noisy data. Multistyle training (combining several noise conditions and reverberation conditions) can yield good trade-offs, even if matching room characteristics in the training and testing conditions gives best performance. In the specific case of convolution, we concluded that synthesizing additional impulse responses via the high-level modeling approach described in [2] avoids overtraining when only one IR is available for the room.

6. Acknowledgements

This work was partly funded by the European Commission's Information Societies Technologies (IST) Programme.

7. References

[1] SPEECON – Speech-Driven Interfaces for Consumer Devices, `http://www.speecon.com/`

[2] O. Gedge, C. Couvreur, K. Linhard, S. Shammass, and A. Moyal, "Database adaptation for speech recognition in cross-environmental conditions," in *Proc. LREC2002*, Las Palmas, Canary Islands, Spain, May 2002.

[3] L. Couvreur, C. Couvreur, and C. Ris, "A corpus-based approach for robust ASR in reverberant environments," in *Proc. ICSLP 2000*, Beijing, China, Oct. 2000, vol. I, pp. 397–400.

[4] F. Diehl, V. Fischer, A. Kiessling, K. Marasek, *Specification of Databases – Specification of recording scenarios*, SPEECON Deliverable D212, March 2001.

[5] S. Van Gerven and F. Xie, "A Comparative Study of Speech Detection Methods", in *Proceedings EUROSPEECH '97*, Rhodes, Greece, Sept. 1997, pp. 1095–1098.

[6] The HTK Toolkit, `http://htk.eng.cam.ac.uk/`