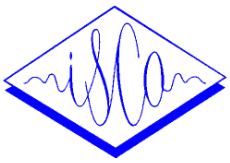

6th International Conference on Spoken Language Processing (ICSLP 2000)
Beijing, China
October 16-20, 2000

A CORPUS-BASED APPROACH FOR ROBUST ASR IN REVERBERANT ENVIRONMENTS

Laurent Couvreur[†] *Christophe Couvreur*[‡] *Christophe Ris*[†]

[†]Faculté Polytechnique de Mons [‡]Lernout & Hauspie Speech Products
e-mail: {lcouv,ris}@tcts.fpms.ac.be | christophe.couvreur@lhs.be

ABSTRACT

In this paper, we discuss the use of artificial room reverberation to increase the performance of automatic speech recognition (ASR) systems in reverberant enclosures. Our approach consists in training acoustic models on artificially reverberated speech material. In order to obtain the desired reverberated speech training database, we propose to use a reverberating filter whose impulse response is designed to match two high-level acoustic properties of the target reverberant operating environment, namely the early-to-late energy ratio and the reverberation time. Speech recognition experiments in simulated reverberant environments show that recognizers trained on speech reverberated with the proposed method outperform systems trained on clean speech, even when channel normalization methods like CMS and logRASTA-PLP are used. The extension of our approach to multi-style training is also considered.

1. INTRODUCTION

Recognition of distant-talking speech is a promising technology for man-machine interaction. Unfortunately, in many applications the operating enclosure is reverberant and the distance between the speech source and the microphone is higher than the so-called critical distance [6]. That is, most of the acoustic energy reaches the microphone after one or more reflections and the recorded speech signal is highly reverberated. The speech signal is severely distorted by this room reverberation, leading to degraded performance of speech recognizers [7].

Several methods have been proposed to cope with room reverberation in speech recognition applications. In some methods, speech is enhanced prior to the extraction of the usual acoustic features [9, 7]. In other methods, robust acoustic features are computed directly from the reverberated speech via channel normalization techniques such as cepstral mean subtraction (CMS) [3] or RASTA-like algorithms [4, 5]. Unfortunately, these methods fail to yield satisfying results on highly reverberated speech.

The discrepancy between the training conditions (anechoic speech) and the testing conditions (reverberated speech) accounts for the poor performance of speech recognition in reverberant environments. Thus, one can suggest to train the recognizer on reverberated speech material rather than on anechoic speech material. Ideally, a training database should be collected every time the system has to be deployed in specific reverberant conditions. This approach is obviously not practical. An alternative may be simulating reverberation in order to obtain adequately reverberated training material from an existing clean speech database. To do so, the anechoic speech database can be convolved with an acoustic impulse response measured in the target reverberant environment [7]. However, this approach is problematic because the acoustic impulse response is highly dependent on the geometric and acoustic characteristics of the room, on the source and microphone locations, on the air temperature and humidity, etc [6]. Moreover, reliable measurement of an acoustic impulse response is not straightforward. Thus, it is difficult to guarantee that the measured acoustic impulse response matches perfectly the acoustic impulse response of the target reverberant environment. In practice, this method gives disappointing results.

In this communication, we propose to use a "randomized" reverberating filter instead of a measured acoustic impulse response to obtain the reverberated speech training database. The impulse response of this reverberating filter is designed to match two high-level, perceptually meaningful, acoustic properties of the target reverberant environment, namely the early-to-late energy ratio and the reverberation time.

The paper is organized as follows. In the next section, we describe the proposed method for artificially reverberating speech material. The efficiency of our approach is then assessed by connected digit recognition experiments in reverberant conditions. The experimental set-up is briefly described in section 3 and results are reported in section 4. Conclusions are drawn in Section 5.

2. ARTIFICIAL REVERBERATION

We assume that the effect of room reverberation for a speech recognizer is better characterized by high-level acoustic properties rather than by the fine temporal details of a complete acoustic impulse response. More specifically, we assume that the early-to-late energy ratio G and the reverberation time T_{60} are sufficient to specify room reverberation conditions [6]. The early-to-late energy ratio G is defined as the steady-state ratio between the direct and reverberated sound energies and is expressed in dB. The reverberation time T_{60} is defined as the time interval expressed in seconds in which the sound energy in the room reaches one millionth of its initial value (-60dB) once a sound source is interrupted. We further assume that T_{60} is frequency independent. Under the diffuse sound field assumption, these parameters can be computed easily using the well-known equations of Sabine [6]. Given the geometric and acoustic properties of a reverberant test enclosure, we have

$$G = 10 \times \log_{10} - \frac{S \times D \times \ln(1-\bar{\alpha})}{16 \times \pi \times (1-\bar{\alpha}) \times r^2} \quad (1)$$

$$T_{60} = \frac{\ln 10^6 \times 4 \times V}{c \times \bar{\alpha} \times S} \quad (2)$$

where the parameters S, V, c, $\bar{\alpha}$, D, and r denote the wall surface, the room volume, the speed of sound, the mean wall absorption coefficient, the directivity factor, and the source–microphone distance, respectively. The mean wall absorption coefficient $\bar{\alpha}$ is computed as $1/S \sum_i \alpha_i S_i$ with α_i and S_i standing for the absorption coefficient and the surface of wall i. If the source and the microphone are omnidirectional, the directivity factor D reduces to 1. Note that T_{60} and G can be easily measured in practice: T_{60}

This work was supported in part by a F.R.I.A grant (Fonds pour la formation à la Recherche dans l'Industrie et l'Agriculture, Belgium).

This work was also partly supported by the European LTR Esprit project RESPITE.

can be obtained by applying the interrupted noise method [12], $\bar{\alpha}$ can be derived from V and S via (2), and G can be estimated via (1) if the source–microphone distance r is known.

For our application, a clean database has to be artificially reverberated by computer means. Artificial reverberation can be rendered by convolving clean speech with a finite-impulse-response (FIR) filter. We propose to design the filter to match some desired room reverberation conditions represented by the high-level parameters G and T_{60}. That is, the filter order and the tap amplitudes must be chosen to match the G and T_{60} parameters. Such a reverberating filter can be obtained simply by modulating the envelope of a random sequence with a decreasing exponential function as follows [8, 2]:

1. Generate a Gaussian white noise random sequence $h_1[n]$, $0 \leq n < L$, where the sequence length L is set equal to the integer part of $T_{60} \times F_s$, with F_s denoting the sampling frequency;
2. Decimate the sequence keeping only the taps with the highest amplitudes,
$$h_2[n] = \begin{cases} h_1[n] & \text{if } h_1[n] > \lambda, \\ 0 & \text{otherwise} \end{cases} \quad (3)$$
for some threshold λ;
3. Modulate the sequence with a decaying exponential,
$$h_3[n] = h_2[n] \times \sqrt{e^{-kn}} \quad (4)$$
where the damping constant k is related to the reverberation time T_{60} by the relation: $k = \ln 10^6/(T_{60} \times F_s)$;
4. Scale the direct and early component taps to match the desired early-to-late energy ratio G given some separation time τ between early and late acoustic energies,
$$h[n] = \begin{cases} \sqrt{\gamma} \times h_3[n] & \text{if } n \leq \tau, \\ h_3[n] & \text{otherwise} \end{cases} \quad (5)$$
with the scaling factor γ computed as
$$\gamma = 10^{G/10} \times \sum_{n>\tau} h_3^2[n] \Big/ \sum_{n \leq \tau} h_3^2[n] \quad (6)$$

In practice, the impulse response $h[n]$ is recomputed several times during the reverberation process of a clean speech database. This randomization serves to model the effects of possible source and microphone movements, temperature and humidity changes, etc, within the target reverberant environment.

3. EXPERIMENTAL SET-UP

To assess the performance of our approach, connected digit recognition experiments are performed in simulated reverberant environments. In this section, we briefly describe the experimental framework used to perform these speech recognition experiments.

3.1. Speech Material and Recognizer Description

Speech material used in this work comes from the TIDIGITS database [11] and consists of 12548 sequences of digits, pronounced by 164 speakers (56 men / 57 women / 25 boys / 26 girls). This speech corpus is divided into a training set (12048 utterances) and a test set (500 utterances). Recognition experiments are performed with a hybrid system based on the hidden Markov model / multilayer perceptron (HMM/MLP) paradigm. That is, a MLP is used for the so-called "acoustic modeling" stage, i.e.

Table 1: Various reverberant conditions (the T_{60} column gives the corresponding reverberation times computed with Sabine's formula (2)).

Reverberation	Absorption Coefficient		T_{60} [s]
	Walls	Floor/Ceiling	
R1	0.4	0.6	0.443
R2	0.4	0.4	0.542
R3	0.3	0.5	0.557
R4	0.3	0.3	0.722
R5	0.2	0.4	0.751
R6	0.2	0.2	1.084
R7	0.1	0.3	1.147
R8	0.1	0.1	2.167

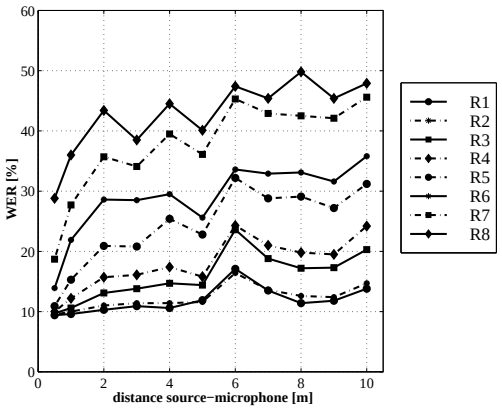

Figure 1: Word error rate as a function of wall absorption coefficients (see Table 1) and source–microphone distance.

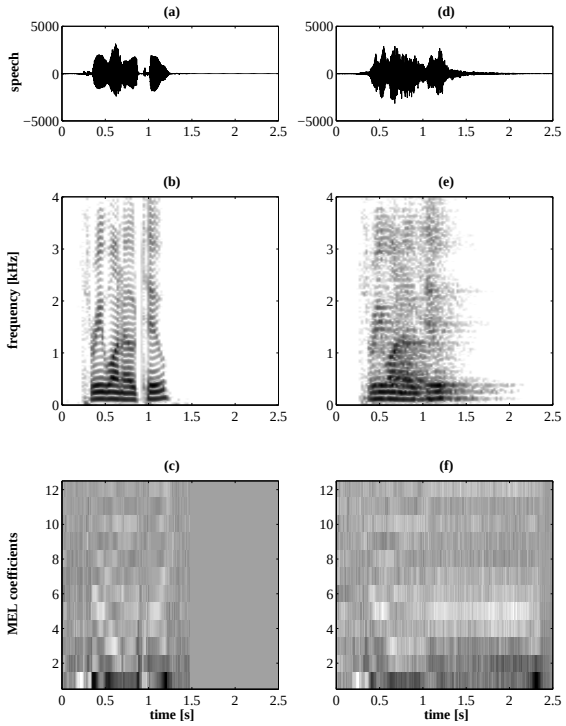

Figure 2: (a) Waveform, (b) spectrogram and (c) MFCC's of clean speech utterance "3201"; (d) waveform, (e) spectrogram and (f) MFCC's of the corresponding reverberated speech utterance.

Table 2: Performance for anechoic speech of baseline recognizers with various front-ends.

Front-end	WER [%]	SUB/DEL/INS [%]
MFCC	4.7	0.7/1.8/2.2
MFCC-CMS	4.9	0.9/1.8/2.2
logRASTA-PLP	5.2	1.3/1.7/2.2

phone probabilities are estimated by the MLP. The recognizer is fed by a front-end producing 12 acoustic coefficients computed over a 25ms analysis window every 12.5ms. Various front-ends are used in this work: Mel-warped frequency cepstral coefficients (MFCC) without cepstral mean subtraction (CMS), MFCC with CMS, and logRASTA-PLP coefficients. Speech decoding is done by Viterbi search, restricted by a wordpair grammar. The complete system is developed with the STRUT toolkit [10].

3.2. Room Acoustic Simulation

For simplicity, our approach is tested in a simulated reverberant environment. An acoustic impulse response can be precisely and efficiently computed by the Image Method [1] for every source–microphone location within a rectangular room whose wall absorption coefficients are known and assumed to be frequency independent and constant over each wall surface. The impulse response can then be convolved with the speech utterances of the test set to provide the reverberated speech material used for the recognition experiments. Several impulse responses are used during the room simulation process so as to take into account possible movements of the speech source (the speaker's mouth) around a reference position.

We want to stress that this room acoustic simulation method is based on a low-level *physical* model of reverberation that requires complete knowledge of the room geometry and of its acoustic characteristics, whereas the artificial reverberation method proposed in Section 2 is based on a high-level *perceptual* model of the reverberation that requires only the two general parameters G and T_{60}.

4. EXPERIMENTAL RESULTS

4.1. Baseline System

First, we trained acoustic models on the clean training set for the three front-ends MFCC, MFCC-CMS and logRASTA-PLP. We then used the resulting MLP's to recognize the clean test set. Table 2 gives the results of these baseline systems in terms of the word error rate (WER) [%], i.e. the sum of the substitution error rate (SUB) [%], the deletion error rate (DEL) [%] and the insertion error rate (INS) [%]. As expected, all these baseline recognizers achieve satisfying results on recognizing anechoic speech. Figure 1 shows the WER's of the MFCC baseline recognizer as a function of the source–microphone distance within a 12m long $\times$ 8m wide $\times$ 6m high test room and for various realistic wall absorption coefficients (see Table 1). The impact of room reverberation on speech recognition can be clearly observed when the test environment (simulated by the method of Section 3.2) becomes more and more reverberant. Figure 2 gives an example of a clean speech utterance and its reverberated version for reverberant conditions R_7 (see Table 1), and the source and the microphone located respectively at point (3m, 4m, 2m) and at point (9m, 4m, 2m), given an origin at one lower corner of the room. This figure also shows the smearing effect of room reverberation on the speech spectrogram and the severe distortion of the corresponding MFCC's. Similar plots can be obtained for the two other front-ends. This explains why the performances of the baseline recognizers degrade severely when the test set becomes reverberant.

4.2. Artificially Reverberated System

Next, we trained acoustic models on artificially reverberated speech material. The reverberated training set was obtained by applying the method described in Section 2. The reverberating filter parameters were computed according to (1)–(2) to match the same simulated reverberant test environment as that of Figure 2: $G = -12.22$dB and $T_{60} = 1.14$s. Table 3 compares the performance of recognizers trained either with the clean training set or with the artificially reverberated training set, and tested on speech reverberated using the room simulation of Section 3.2. We observe that usual channel normalization techniques like CMS or RASTA are inefficient for highly reverberated speech. The reason why these frame-based front-ends do not perform well in handling severe reverberation, is that the duration of the acoustic impulse response is higher than the analysis window length (25ms). Therefore, the reverberation effect can not be approximated by multiplicative noise in the frequency domain within each analysis window. Table 3 also shows that recognizers trained on reverberated speech outperform the other systems. The MFCC recognizer trained with speech reverberated by the proposed "random reverberator" filter can even do better than the MFCC recognizer trained with speech reverberated by the "true" acoustic impulse response corresponding to the reverberant test environment. This can be explained by recalling that several acoustic impulse responses are used to simulate the reverberant test set in order to model small movements of the speaker's head whereas a fixed impulse response is used for the "true impulse response" training set. It leads to over-specific adaptation of the recognizer to the "true" impulse response in the latter case.

The numerous INS errors can be reduced by introducing a *word-entrance-penalty* (WEP) in the Viterbi decoding algorithm [10]. This parameter is tuned to balance DEL and INS errors. Tuning the WEP brings some improvement but the systems trained on reverberated speech still yield the best results (see Table 3). Note that tuning of the WEP is performed on the test set, which provides a lower bound on achievable performance.

In real situation, the room description will usually not be available and the parameters G and T_{60} will have to be measured experimentally. Using the interrupted noise method [12] and (1)–(2) with our simulated reverberant test room, we found $T_{60} = 1.09$s and $G = -11.98$dB, which shows that G and T_{60} can be reliably estimated ($\Delta G = 2.0\%$ and $\Delta T_{60} = -4.4\%$) if they can not be computed from a description of the room. Besides, previous experiments have shown that reasonable deviations of the parameters with respect to the ideal ones can be tolerated [2].

4.3. Multi-Style Training System

Different approaches may be considered to extend our method to practical applications. Though training a recognizer for every reverberant environment is conceivable, this approach remains time-consuming. The last remark of the previous section suggests rather to build a library of recognizers for various reverberant conditions. The recognizer "closest" to the operating reverberant conditions is then picked out of the library. Alternatively, one can propose a *multi-style* training approach: one recognizer is trained on a database including different levels of reverberation, and is used as a universal recognizer for every reverberant environment. Table 4 reports cross-performance of MFCC recognizers trained on different reverberant environments. Four environments were considered: CLEAN (no reverb), LOW ($G = -7.89$dB and $T_{60} = 0.25$s), MEDIUM ($G = -10.08$dB and $T_{60} = 0.68$s) and HIGH ($G = -12.22$dB and $T_{60} = 1.14$s). They served to

Table 3: Performance for reverberant speech of recognizers trained either on clean speech or on reverberated speech, with or without optimal *word-entrance-penalty* in the decoding process.

Front-end	Training set	WER (SUB/DEL/INS) [%]	WER (SUB/DEL/INS) [%]	Optimal WEP
MFCC	clean	45.2 (17.7/2.9/24.6)	33.5 (16.3/8.5/8.7)	WEP = 8.5
MFCC-CMS	clean	44.7 (14.9/2.6/27.2)	28.9 (13.0/7.9/8.0)	WEP = 13.0
logRASTA-PLP	clean	32.9 (13.5/8.1/11.3)	30.4 (12.8/9.1/8.5)	WEP = 0.5
MFCC	random reverb	13.0 (2.2/0.7/10.1)	6.9 (2.9/2.0/2.0)	WEP = 17.5
MFCC-CMS	random reverb	16.3 (3.4/0.6/12.3)	7.9 (3.6/2.1/2.2)	WEP = 18.0
logRASTA-PLP	random reverb	20.9 (4.8/0.6/15.5)	10.0 (5.0/2.5/2.5)	WEP = 18.5
MFCC	true reverb	13.9 (2.3/0.8/10.8)	7.3 (2.3/2.5/2.5)	WEP = 16.5

Table 4: Performance WER(SUB/DEL/INS) [%] of various acoustic models for several reverberant environments. Acoustic features are MFCC. Speech decoding is performed with *word-entrance-penalty*.

Test set	Training set				
	CLEAN	LOW	MEDIUM	HIGH	ALL
CLEAN	**4.7 (0.7/1.8/2.2)**	5.8 (2.0/1.9/1.9)	8.5 (4.1/2.2/2.2)	11.7(6.2/2.8/2.7)	**4.9 (0.8/2.0/2.1)**
	WEP=30.0	WEP=24.0	WEP=6.0	WEP=4.0	**WEP=31.5**
LOW	6.9 (2.8/2.0/2.1)	**5.5 (1.6/1.9/2.0)**	9.6 (5.6/2.0/2.0)	10.5 (6.0/2.3/2.2)	**5.7 (1.3/2.2/2.2)**
	WEP=22.5	**WEP=26.0**	WEP=17.0	WEP=13.5	**WEP=29.0**
MEDIUM	11.2 (3.7/3.7/3.8)	7.0 (1.9/2.5/2.6)	**6.0 (2.4/1.8/1.8)**	7.0 (3.6/1.7/1.7)	**6.4 (2.3/2.0/2.1)**
	WEP=17.0	WEP=24.0	**WEP=20.5**	WEP=16.5	**WEP=23.0**
HIGH	28.1 (9.9/9.1/9.1)	22.1 (5.9/8.1/8.1)	15.7 (3.4/6.1/6.2)	**6.9 (2.9/2.0/2.0)**	**9.4 (3.6/2.9/2.9)**
	WEP=10.5	WEP=19.0	WEP=34.0	**WEP=17.5**	**WEP=20.0**

generate both a training set and a test set, the former by "random" reverberation and the latter by room simulation. The HIGH reverberant environment corresponds to the environment which has been used so far. Clearly, highest scores are reached when training and testing sets match. The last column shows the performance of the ALL recognizer which is obtained by multi-style training on the CLEAN, LOW, MEDIUM and HIGH training sets. Though the ALL recognizer gets slightly worse scores than the recognizer matching the test set, it performs satisfactorily on every test sets.

5. SUMMARY AND CONCLUDING REMARKS

In this paper, we discussed the use of artificial room reverberation to improve speech recognition in reverberant enclosures. Our approach consists in training acoustic models on reverberated speech material. In order to avoid having to collect a reverberated database or to measure acoustic impulse responses, we proposed to obtain the reverberated database by processing a clean speech corpus with a "random" reverberating filter whose impulse response is obtained by shaping the envelope of a Gaussian white noise random sequence. This reverberating filter is designed to match two high-level, perceptually meaningful, acoustic properties of the desired reverberant operating environment, i.e. the early-to-late energy ratio and the reverberation time.

Connected digit recognition experiments were performed in simulated reverberant environments. Experimental results showed that recognizers trained on reverberated speech outperform systems trained on clean speech, even when conventional channel normalization methods like CMS and RASTA are used. We also demonstrated the usefulness of a multi-style training approach.

Similar results were observed for real reverberated speech where the method has been found to improve the performance significantly, provided that the parameters of the reverberating filter are close to the true ones.

6. REFERENCES

[1] J. B. Allen and D. A. Berkley, "Image Method for Efficiently Simulating Small-Room Acoustics", *Journal of the Acoustical Society of America*, vol. 65, no. 4, pp. 943–950, Apr. 1979.

[2] L. Couvreur and C. Couvreur, "On the Use of Artificial Reverberation for ASR in highly reverberant Environments", *Proc. SPS'2000*, Hilvarenbeek, The Netherlands, Mar. 2000.

[3] S. Furui, "Cepstral Analysis Technique for Automatic Speaker Verification", *IEEE Trans. on Acoustics Speech and Signal Processing*, vol. 29, no. 2, pp. 254–272, Apr. 1981.

[4] H. Hermansky and N. Morgan, "RASTA Processing of Speech", *IEEE Trans. Speech and Audio Processing*, vol. 2, no. 4, pp. 578–589, Oct. 1994.

[5] B. E. Kingsbury and N. Morgan, "Recognizing Reverberant Speech with RASTA-PLP", *Proc. ICASSP'97*, vol. 2, pp. 1259–1262, Oct. 1997.

[6] H. Kuttruff, *Room Acoustics*, Applied Science Publishers, 2nd edition, 1979.

[7] M. Matassoni, M. Omologo and D. Giuliani, "Hands-Free Speech Recognition Using Filtered Clean Corpus and Incremental HMM Adaptation", *Proc. ICASSP'2000*, Jun. 2000.

[8] J. Moorer, "About this Reverberation Business", *Computer Music Journal*, vol. 3, no. 2, pp. 13–18, 1979.

[9] S. Nakamura and K. Shikano, "Room Acoustics and Reverberation: Impact on Hands-Free Recognition", *Proc. EUROSPEECH'97*, vol. 5, pp. 2419–2422, Sep. 1997.

[10] *STRUT* - MULTITEL-TCTS Lab, Faculté Polytechnique de Mons, Belgium, `http://tcts.fpms.ac.be/asr/strut.html`.

[11] "TIDIGITS speech corpus", Linguistic Data Consortium, University of Pennsylvania, Philadelphia, PA, USA, 1993.

[12] M. Vorländer and H. Bietz, "Comparison of Methods for Measuring Reverberation Time", *Acustica*, vol. 80, pp. 205–215, 1994.