A STUDY OF SPEECH DISTORTION CONDITIONS IN REAL SCENARIOS FOR SPEECH PROCESSING APPLICATIONS

Dayana Ribas[1], *Emmanuel Vincent*[2], *José Ramón Calvo*[1]

[1]Advanced Technologies Application Center (CENATAV), Habana, Cuba
[2]Inria, Villers-lès-Nancy, F-54600, France

{dribas,jcalvo}@cenatav.co.cu, emmanuel.vincent@inria.fr

ABSTRACT

The growing demand for robust speech processing applications able to operate in adverse scenarios calls for new evaluation protocols and datasets beyond artificial laboratory conditions. The characteristics of real data for a given scenario are rarely discussed in the literature. As a result, methods are often tested based on the author expertise and not always in scenarios with actual practical value. This paper aims to open this discussion by identifying some of the main problems with data simulation or collection procedures used so far and summarizing the important characteristics of real scenarios to be taken into account, including the properties of reverberation, noise and Lombard effect. At last, we provide some preliminary guidelines towards designing experimental setup and speech recognition results for proposal validation.

Index Terms— real applications, robust speech processing, reverberation, noise, Lombard effect

1. INTRODUCTION

One of the challenges currently faced in many areas of speech processing, including speech enhancement, automatic speech recognition (ASR), speaker identification and verification, keyword spotting, and computational paralinguistics, is the migration of laboratory results to real application scenarios [1–4]. The acoustic variability in these scenarios raises several robustness problems due to environmental noise, reverberation, and source or sensor movement, among others. The best way to ensure that a scientific result has practical value is to evaluate it in the targeted scenarios of use, e.g., [5–9]. However, this is rarely feasible in the early stages of research. Researchers have therefore resorted to evaluating their methods with recorded or simulated data in controlled scenarios.

It is usual that simulated datasets are acoustically unrealistic. For example, it is not uncommon to mix speech signals together without noise or reverberation, e.g., [10], or to simulate noise alone, e.g., [11, 12], or reverberation alone, e.g., [13, 14], instead of considering them in combination as they appear in real indoor environments. The noise signals may also be overly simplistic, e.g., Gaussian random noise [15, 16] or individually recorded noises added together [17–19], instead of real recorded noise scenes. The noise and reverberation levels are frequently selected independently from each other without a specific application in mind, e.g., [20, 21], therefore some of them might never happen in real life. Furthermore, they are often selected within a discrete set of values, e.g., [11, 14, 22–24] or a narrow range of values, e.g., [25], which does not match the actual distribution of levels observed in real life and artificially advantages learning-based methods which may overfit those levels. Even when the distortion levels are realistic, there may still exist some acoustic mismatch, due to recording speech in a different place than noise and reverberation, e.g., [26, 27]. At last, the Lombard effect is often not considered, e.g., [28, 29].

Although real data are exempt from such criticisms, they are still not always ecologically realistic because they were not collected in the real conditions of use. Depending whether the considered artificial scenario is simpler or more difficult than the real one, this can result in selecting methods that will perform poorly in real scenarios or whose high computational cost is actually not required.

The information about which conditions are required for a dataset to be realistic and which ones are actually important for the evaluation of a certain task is sparsely found in the literature. We have identified above some of the main problems with simulated or real data used for robust speech processing experiments so far. In the rest of this paper, we list the characteristics of real scenarios to be taken into account in Section 2, and summarize in Section 3 the important properties of reverberation, noise and Lombard effect, and validate them on a real robust ASR dataset. These properties were established in different papers in the acoustics community and are still little known in the speech community. Then, we propose an approach for defining the scenario distortion pattern and a procedure for configuring the experimental setup regarding the target application in Section 4, followed by some ASR results for proposal validation. We conclude in Section 5.

2. CHARACTERIZATION OF REAL SCENARIOS

2.1. Interface and speaker distance

Many factors influence the speech distortion conditions in real scenarios. Among them, the recording device used for signal acquisition defines the typical distance to the speaker (D), which influences the distortion level [30]. For large D, multi-microphone interfaces are generally used. The number of microphones influences in turn the choice and the performance of the distortion compensation method.

As it can be seen in Table 1, D varies from one application to another. When talking over the phone or recording TV/radio speakers in a studio, the speaker's mouth is very close to the recording device. For access control as well as hands-free call/control devices such as tablets or in-car GPS, D is larger but it still lies in a small interval. By contrast, in some forensics applications, the speaker is not aware of the position of the recording device, so D is quite variable. Similarly, certain applications such as audioconferencing and voice-controlled home automation rely on voice call/control devices which can both be held close to the user's mouth or operate at distance, resulting in a large interval of distances.

Table 1. *Characterization of interfaces used in speech processing applications by number of microphones (Ch.), possibility of sensor movement (Move.), and distance D to the speaker.*

#	Interface	Ch.	Move.	D (cm)
1	Access control devices	1	No	$\approx$ 30–70
2	Cell/landline phone	1-2	Yes/No	$\approx$ 3–15
3	Computer/tablet/PDA	2	Yes	$\approx$ 30–50
4	Hidden microphone	1-N	Yes	$\approx$ 10–1000
5	Voice call/control device	1-N	Yes	$\approx$ 5-500
6	Studio microphone	1-N	No	$\approx$ 3-25

2.2. Scenario location

The scenario location, i.e., indoor or outdoor, helps to define the possible type of distortion in the scene. Outdoor scenarios are only affected by environmental noise. For indoor scenarios, the Room Impulse Response (RIR) characteristics should be considered in addition. Table 2 shows examples of scenario locations for different speech processing applications.

Table 2. *Scenario location for speech processing applications.*

Applications	Scenario location	Interface (Table 1)
Access control to physical facility, network, or website	indoor/outdoor	1, 3
Voice service (e.g., banking)	indoor/outdoor	2, 3
Forensics	indoor/outdoor	2, 3, 4
Home parole[1]	indoor	2
Home automation, in-car GPS	indoor	5
Intelligent answering machine	indoor	2
TV/radio stream	indoor/outdoor	6
Hearing aid	indoor/outdoor	5
Meeting/lecture transcription	indoor	3, 5

2.3. Distortion type

The types of noises should be defined according to the scenario environment in the specific application. See some examples in Table 3. Real noise scenes are usually mixtures of different noises with diffuse spatial distributions and complex time dependencies, which remain hard to simulate today.

In indoor scenarios, the RIR should be considered. It consists of: direct sound, early echoes, and reverberation (late reflections). Reverberation spreads the speech spectrum over multiple time frames [31]. Dereverberation algorithms typically aim to cancel this spreading effect. Early echoes result in a channel effect instead [31]. They color the speech spectrum in the current time frame and require channel-robust algorithms based on, e.g., cepstral mean normalization or factor analysis.

3. PROPERTIES OF DIFFERENT TYPES OF DISTORTION

So far, we have grossly characterized the scenarios in terms of speaker distance, scenario location, and distortion type. This is illustrated in Fig. 1.

[1]Remote verification of people under parole, i.e., required to remain at home by law.

Table 3. *Examples of environmental noises in common speech processing application scenarios.*

Scenario	Environmental Noises
Restaurant/Cafeteria	Cutlery, voices, music
Home	Voices, TV, electrical appliances
Conference/Meeting	Voices, chairs, air conditioning
Train/Metro station	Train, voices, footsteps, music
Airport	Traffic, music, air conditioning
Car/Bus	Cockpit noise, engine, wind, traffic
Office	Phone ring, air conditioning, typing
Street/Square/Park	Traffic, footsteps, voices, wind
TV/Radio stream	Music, applause

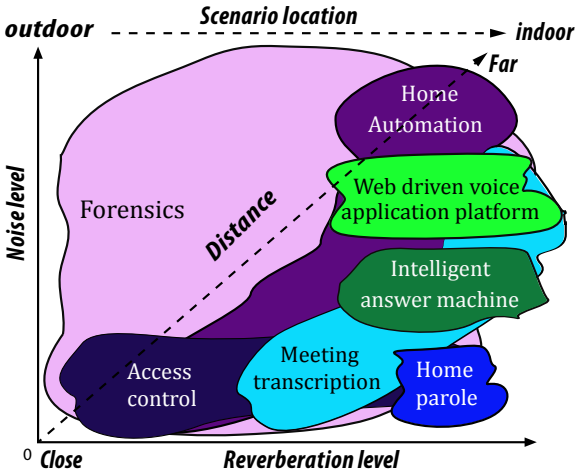

Fig. 1. *Distortion types in speech processing applications.*

We now analyze these distortions in more detail and quantify some of their properties. Besides environmental distortions, such as reverberation and noise, there are additional speaker variability factors that can be significant in some specific application scenarios, e.g. emotion, illness, stress, aging, etc. Among them we consider the Lombard effect, that is the phenomenon by which speakers tend to raise their voices when speaking in a noisy environment. Others intraspeaker variabilities are hard to simulate over existing speech signals, so this should be considered at the time of corpus selection.

3.1. Reverberation properties

Reverberation is characterized by three main properties which depend on the acoustic features of the room and the speaker/microphone positions. One of the most widely used is the time taken for the reverberant energy to decay by a certain amount (usually 60 dB) once the sound has been abruptly shut off. This is the well-known Reverberation Time (RT_{60}), which only depends of the acoustic features of the room, so it is constant at any location in the room [32]. Different scenarios translate into different RT_{60}. The relationship between $RT_{60}(\mathrm{s})$, the room volume $V(\mathrm{m}^3)$, the amount of air absorption $A(\mathrm{m}^2)$, the total room surface $S(\mathrm{m}^2)$ and the average surface absorption coefficient α is approximated by Eyring's formula [33]:

$$RT_{60} = 0.161 \frac{V}{A - S\log(1-\alpha)}. \quad (1)$$

Previous works have reported RT_{60} values for specific scenarios (see Table 4). Furthermore, it was found that RT_{60} is approximately 0.1 s larger for unoccupied rooms compared to occupied rooms, showing that humans absorb the sound energy, reducing the reverberation [34, 35]. However, this is practically meaningful only in crowded scenarios [36], where the RT_{60} should be computed from people's absorption, which mainly depends on the clothes [37].

Table 4. *RT_{60} reported in the literature for different scenarios (unoccupied rooms).*

Scenario	RT_{60}**(s)**
In-car [38]	0.05
Studio booth [39]	0.12
Living room [38, 40, 41]	0.44–0.74
Bedroom [40, 41]	0.39–0.68
Bathroom [40, 41]	0.41–0.75
Corridor [40]	0.60
Kitchen [40, 41]	0.41–0.83
Restaurant [37]	0.5–1.5
Elevator hall [38]	0.75
Building lobby [35]	0.65
Office [35, 38, 39]	0.25–0.43
Meeting room [35, 38, 39, 42]	0.23–0.70
Classroom [43–45]	0.20–1.27
Lecture room [35, 39]	0.64–1.25

Another property of reverberation is the level of reverberant energy vs. direct sound (or vs. direct sound and early echoes) in the recorded speech signal. The Direct-to-Reverberant Ratio (DRR) in decibels (dB) depends on the acoustic features of the room and on D. For omnidirectional microphones, it is approximately given by [32]

$$DRR = 10\log_{10}\frac{S\alpha}{16\pi D^2(1-\alpha)}. \quad (2)$$

Fig. 2 shows the relation between the DRR, RT_{60}, and D. The Early to Late Ratio (ELR), also called clarity index or C_{50}, is an alternative metric that is widely employed. It is computed as the energy of the first 50 ms of the RIR divided by the remaining energy [32].

One last property is the level of speech coloration induced by early echoes. This could be quantified in a similar way by the Direct-to-Early Ratio (DER), that is the energy of direct sound divided by the remaining energy in the first 50 ms. Early echoes are a serious concern in many scenarios, for instance when the microphone and/or the speaker is close to a table or a window. It is therefore essential to test algorithms on data with different ELRs and DERs as relevant for the targeted scenario.

In the remainder of this paper, we will assume for simplicity that the targeted scenario does not involve significant early echoes. In that case, the channel effect due to early echoes is limited and specifying the ELR becomes essentially equivalent to specifying the DRR.

3.2. Noise level

Classically, the amount of noise in a given scene can be measured by the Signal-to-Noise Ratio (SNR) computed from the noise and speech levels (L_N, L_S). Previous works have studied the noise level in different scenarios based on the loudness profiles of recordings expressed in dB [46–48], considering type *A* frequency weighting [30]. The speech level exhibits variations [49] due to the speaker identity,

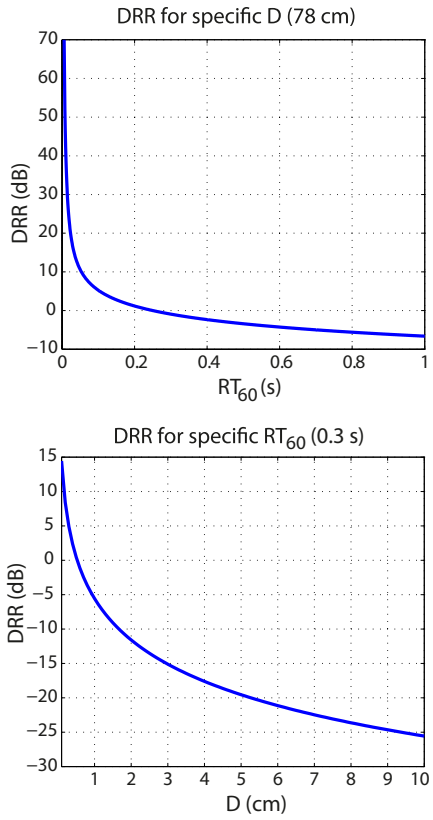

Fig. 2. *Relationship of DRR with RT_{60} and D (fixed room size of $4.45 \times 3.55 \times 2.50$ m).*

the verbal contents, the speech modality (isolated word, short command, read speech, spontaneous speech, dialogue), and the Lombard effect.

3.3. Lombard effect

The Lombard effect arises when producing speech in the presence of noise [50]. The vocal effort, and as a result the speech intensity, are increased in order to achieve more effective communication. It also induces other variations such as increasing the pitch, the first formant, and high-frequency components in general, which is a consequence of the fall of spectral tilt. It also results in time variations, increasing the duration of vowels and decreasing that of fricatives and stops, and moves the energy to the medium frequency range in vowels and to higher frequencies for stops and fricatives. Previous work has shown that these variations are dependent on noise type and levels. The intelligibility is increased up to certain noise level until the speaker starts shouting and the intelligibility abruptly decreases. Due to the range of acoustic-phonetic modifications implied by Lombard speech, automatic speech and speaker recognition performance also decrease unless suitable treatment is applied [51, 52].

Coming back to the impact of Lombard effect on speech intensity, the ISO-9921 standard [53] provides a piecewise affine model for the range of speech levels $L_{S,A,1m}(1)$ observed at 1 m when talking to a person at $D = 1$ m as a function of the noise level $L_{N,A}$ (see Fig. 3). This model is valid in the case when the propagation of sound from the speaker's mouth is not obstructed, and the level and spectrum of noise are identical at the speaker and the listener. The

latter condition holds for small D. It also often holds for large D, depending on the position of the speaker relative to the microphone and the noise sources. Considering these results, the rate of increase of speech level as a function of the noise level has been quantified by defining the Lombard slope C_{Lomb}. This quantity varies from 0.55 to 0.64 dB/dB for $45 < L_{N,A} < 82$ dB in Fig. 3 and it has been reported to span a larger range from 0.5 to 0.7 dB/dB in [49].

A similar model was proposed for the variation of speech level as a function of the distance D to the listener. When talking to a person at a larger (resp. smaller) distance D, the speech level increases (resp. decreases) by $c_{\text{dist}} \times 20\log_{10}(D)$ in the limit of the maximum (resp. minimum) physically achievable speech level. Lazarus [49] reported a value of $c_{\text{dist}} = 1$ dB/dB.

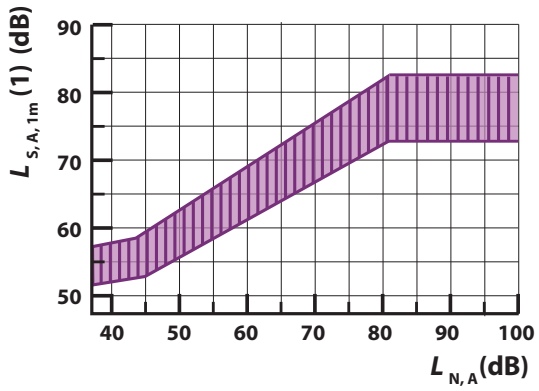

Fig. 3. *Speech level $L_{S,A,1m}(1)$ at 1 m when talking to a person at $D = 1$ m vs. noise level $L_{N,A}$, considering the expected vocal effort. The speech level values for noise levels below and above 82 dB are taken from [53, Fig. A.1] and [53, Table A.1], respectively.*

3.4. Calculation of the SNR

Combining these results, we can compute the range of SNRs for a given target application as follows. Let $L_{N,A}$ be the A-weighted noise level taken from, e.g., [46–48]. The typical range of D depends on the interface employed in this application, as exemplified in Table 1. For $45 < L_{N,A} < 82$ dB, the A-weighted speech level in dB at 1 m when talking to an interface at distance D can be expressed as

$$L_{S,A,1m}(D) = L_{spk} + C_{\text{Lomb}}(L_{N,A} - 45) + c_{\text{dist}} \times 20\log_{10}(D) \quad (3)$$

where L_{spk} is a speaker-dependent value which varies between 53 and 59 dB according to Fig. 3. Recall that this expression holds only in the limit of physically achievable speech levels and under certain conditions explicited in Section 3.3. Denoting by $DL(D)$ the sound level attenuation as a function of distance defined in [49] as

$$DL(D) = 20\log_{10}(D), \quad (4)$$

the speech level at the interface ($L_{S,A,mic}$) is expressed as

$$L_{S,A,mic} = L_{S,A,1m}(D) - 20\log_{10}(D). \quad (5)$$

The A-weighted SNR in dB is then obtained as

$$\begin{aligned} SNR = &= L_{S,A,mic} - L_{N,A} \quad (6)\\ &= L_{spk} - C_{\text{Lomb}} \times 45 \\ &+ (C_{\text{Lomb}} - 1)L_{N,A} \\ &+ (c_{\text{dist}} - 1) \times 20\log_{10}(D). \quad (7) \end{aligned}$$

This means that, in the limit of physically achievable speech levels and under the above conditions, the SNR at the interface decreases by $6(C_{\text{Lomb}} - 1)$ dB per doubling of noise amplitude and $6(c_{\text{dist}} - 1)$ dB per doubling of distance, respectively. Similar expressions can be derived for other noise levels, taking the minimum and maximum physically achievable speech levels into account.

3.5. Validation on the CHiME-3 dataset (real data)

In order to assess how well these equations hold in a practical robust speech processing application, we measure the speech level and the noise level on the real data of the 3rd CHiME Speech Separation and Recognition Challenge (CHiME-3) [9]. This data consists of utterances from the Wall Street Journal (WSJ0) corpus [54] spoken live by twelve US English talkers in four everyday environments: bus, café, pedestrian area, and busy street. Recordings are made by a tablet equipped with an array of six microphones, held at an average distance of 30 cm from the face, and sampled at 16 kHz. The dataset involves 4560 real utterances in total: 1600 for training, 1640 for development, and 1320 for test. The training, development, and test data were recorded in different instances of each environment (e.g., different buses). The start and end time of all utterances are annotated. Besides these noisy data, the CHiME-3 dataset also includes noiseless WSJ0 utterances recorded by the same talkers in a sound proof booth using the same recording hardware.

In order to separate each noisy recording into speech and noise components, speech was captured by a close-talking microphone in parallel with the array. For each noisy utterance, we estimated the time-varying relative impulse responses between the close-talking microphone and the array microphones in the least squares sense. By convolving the estimated impulse responses with the close-talking microphone signal, we obtained estimates of the speech signals at the array microphones. We then derived the noise signals by subtracting the estimated speech signals from the recorded signals. For a more detailed explanation of this procedure, see [9].

Fig. 4 shows the A-weighted levels of speech and noise (averaged over time and over all microphones) for all noisy utterances. The analog-to-digital conversion factor was set such that the average A-weighted level of speech in the booth recordings is equal to the

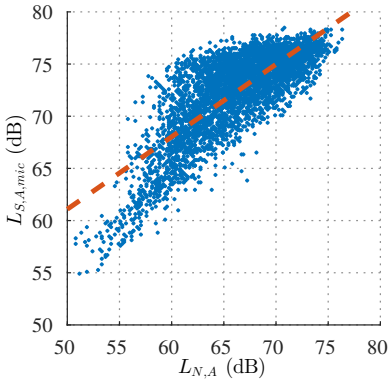

Fig. 4. *Measured speech level at the microphones ($L_{S,A,mic}$) vs. noise level ($L_{N,A}$) for the 4560 real noisy utterances of the CHiME-3 dataset [9]. Sound levels are averaged over all six microphones. The slope of the linear least squares fit (dashed line) is $C_{\text{Lomb}} = 0.69$ dB/dB.*

level of normal speech at a distance of 30 cm, namely 71 dB[2]. The figure confirms that, for a given noise level, a range of speech levels are observed. If we concentrate on denser areas and disregard outliers, the width of that range appears comparable to that in Fig. 3. A linear least squares fit indicates that speech level increases with noise level at an average rate of $C_{\text{Lomb}} = 0.69$ dB/dB, which is consistent with the range of Lombard slopes reported in [49].

4. IMPACT ON DATASET CREATION AND PERFORMANCE RESULTS

4.1. Tentative dataset creation and characterization procedure

Based on the above results, we propose a tentative procedure for creating an adequate experimental setup considering the acoustic characteristics of the potential target application scenarios (see Procedure 1). This proposal seeks to guide the process of collecting real data, or failing that, the generation of simulated data, in order to obtain realistic types and levels of distortion regarding the application. In order to help the definition of the distortion type in step 6, Table 3 shows some examples of noise types for different scenarios. However, only the precise application scenario can actually define the specific noise types in the scene. Therefore the most rigorous approach is to acquire real noise samples in the actual scenario.

Procedure 1: Dataset selection for experimental setup

1 *Application*: Define application of interest
2 *Scenario*: Define the potential scenarios
3 *Interface*: Define the recording interface (Table 1)
4 D: Define the speaker distance (Table 1)
5 *Location*: Identify the scenario location (Table 2)
6 Characterize the distortion type:
a) Reverberation:
if *Location* = *indoor* **then**
Define $RT_{60}(s)$ (Table 4)
$DRR(dB)$: Compute the DRR using (2)
b) Noise:
if *You have access to the target scenario* **then**
Measure $L_{N,A}$ and $L_{S,A,mic}$ (dB)
else
Get $L_{N,A}$ (dB) from a dataset [46–48] and predict $L_{S,A,mic}$ via (5)
7 $SNR(dB)$: Compute the SNR using (6)

Table 5 and Fig. 5 illustrate the use of this procedure for two different applications. Table 5 describes the experimental setup and the range of D, RT_{60}, DRR, and SNR for the dataset creation. These ranges appear to be very different for the two applications. Fig. 5 shows the values of DRR and SNR as a function of D, RT_{60}, $L_{N,A}$ and $L_{S,A,1m}(1)$. It is clear that these values can't be chosen independently, since they strongly correlate with each other.

4.2. Impact on the CHiME-2 Track 1 dataset (simulated data)

In order to illustrate how this dataset creation procedure may change the results reported in the literature on simulated data, we consider

[2]Note that, although this choice affects the absolute speech and noise levels displayed in Fig. 4, it does not affect the measurement of the SNR and the Lombard constant, which are the actual quantities of interest.

Table 5. *Example of application of Procedure 1.*

Step	Fig. 5a	Fig. 5b
Application	Access control	Voice call
Scenario	Office	Kitchen
D (cm) (Tab. 1)	30 – 70	5 – 500
Location	Indoor	Indoor
RT_{60} (s) (Tab. 4)	0.25 – 0.43	0.41 – 0.83
DRR (dB)	-2 – 9	-22 – 22
$L_{N,A}$ (dB) [48]	40 – 60	60 – 90
$L_{S,A,1m}(1)$ (dB) (Fig. 3)	52 – 69	61 – 83
C_{dist} (dB/dB) [49]	1	1
SNR (dB)	1.3 – 27.5	-30.6 – 23.0

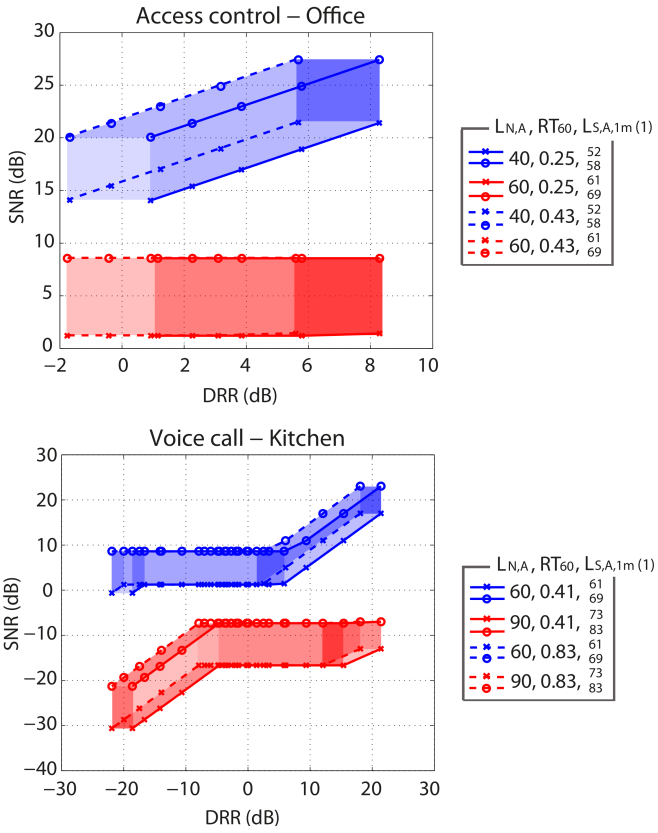

Fig. 5. *Computed SNR vs. DRR in the two scenarios of Table 5 as a function of D. Dots correspond to 30, 40, 50, 60, 70 cm (Office) and 5, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 200, 300, 400, 500 cm (Kitchen). Short distances are on the top right, large distances on the bottom left.*

the task of robust ASR on Track 1 of the 2nd CHiME Speech Separation and Recognition Challenge (CHiME-3) [28]. The considered scenario is that of a speaker uttering short commands recorded by a distant pair of microphones in a domestic environment.

Clean speech signals from 34 speakers are taken from the Grid corpus [55]. They are 6-word sequences of the form <command:4><color:4><prepos.:4><letter:25><digit:10><adverb:4>, with the numbers in brackets indicating the number of choices per word. The task is to recognize the letter and digit tokens. Success is measured by the keyword recognition rate, that is the percentage of cor-

Table 6. *Noise level ($L_{N,A}$), speech level at the microphone ($L_{S,A,mic}$), SNR, and keyword recognition accuracy on Track 1 of the CHiME-2 dataset [28] with/without taking the Lombard effect into account in the data simulation. The subset names are those used to refer to the various noise levels in the original dataset and they do not reflect the actual SNR after taking the Lombard effect into account.*

		Dev						Test					
Subset name		-6dB	-3dB	0dB	3dB	6dB	9dB	-6dB	-3dB	0dB	3dB	6dB	9dB
$L_{N,A}$ (dB)		61.8	58.1	55.1	51.7	48.5	45.3	61.4	57.8	54.8	51.5	48.3	45.0
Without Lombard effect [28]	$L_{S,A,mic}$ (dB)	55.0						55.0					
	SNR (dB)	-5.8	-2.8	0.1	3.2	6.2	9.2	-5.5	-2.6	0.4	3.4	6.5	9.5
	Key. acc. (%)	**49.67**	**57.92**	**67.83**	**73.67**	**80.75**	**82.67**	**49.33**	**58.67**	**67.50**	**75.08**	**78.83**	**82.92**
With Lombard effect	$L_{S,A,mic}$ (dB)	66.1	63.7	61.8	60.1	58.3	56.7	65.6	63.6	61.7	59.8	58.0	56.3
	SNR (dB)	5.4	5.9	7.0	8.2	9.5	11.0	5.2	6.1	7.2	8.3	9.6	10.9
	Key. acc. (%)	**80.67**	**81.25**	**82.58**	**83.83**	**85.83**	**86.00**	**81.75**	**83.50**	**85.33**	**85.83**	**87.92**	**87.00**

rectly recognized tokens. The clean speech signals are convolved with binaural room impulse responses (BRIRs) and mixed with real noise backgrounds recorded using two ear microphones built into a dummy head placed at a fixed position in a living room. About 14 h of noise backgrounds (including concurrent speakers, TV, game console, footsteps, distant noises...) were collected over a period of several days. The BRIRs were measured for various positions covering an area of 20 cm side centered on the position 2 m directly in front of the microphone pair, in order to simulate small movements of the speaker. The dataset is organized into 6 subsets corresponding to different ranges of noise levels. For simplicitly, each subset was named according to the corresponding range of SNRs, that is roughly centered around -6, -3, 0, 3, 6, or 9 dB, respectively. For each of the 6 subsets, 600 noisy development utterances and 600 noisy test utterances sampled at 16 kHz are provided. The same clean speech signals are used for all SNRs but the development and test utterances are different. A separate training set of 500 utterances is also provided for each of the 34 Grid talkers. Each training utterance is mixed at a random SNR among the 6 ranges above.

In the original dataset, the speech signals were rescaled so that their level was fixed (utterance-independent) and matched that of a human speaker at a natural conversational level[3]. For each utterance, the background recordings were scanned until a time interval resulting in the desired SNR range was found. It was observed that the backgrounds at 9 dB are dominated by quasi-stationary ambient sources, while those at -6 dB typically involve nonstationary, sudden sound events. As acknowledged by the authors, this mixing procedure was ecologically more valid than mixing any kind of noise at any SNR, but it did not account for the Lombard effect.

In the following, we modified the dataset by taking the Lombard effect into account. For each utterance, we measured the A-weighted level of the noise signal (averaged over time and over the two microphones), derived the minimum and maximum A-weighted speech levels according to (5) (generalized to noise levels below 45 dB as in Fig. 3), rescaled the speech signal after convolution with the BRIR to a random value within that range, and mixed it with the noise signal. We set the analog-to-digital conversion factor such that the average A-weighted level of speech after convolution is equal to that of normal speech at a distance of 2 m, namely 55 dB. Note that only the speech level has changed compared to the original dataset: the clean speech signal, the BRIR, the noise signal and the noise level corresponding to each utterance of each subset are unchanged.

We performed ASR on the new dataset using the baseline system based on HTK [56] made available by the challenge organizers. The features are standard 39-dimensional Mel-frequency cepstral coefficient (MFCC) vectors (including delta and acceleration coefficients) with cepstral mean normalization. Each of the 51 words in the Grid vocabulary is modeled with a speaker-dependent left-to-right hidden Markov model with Gaussian mixture model densities (GMM-HMM) trained on the noisy training set. The language model is fixed according to the Grid syntax. For more details, see [28].

The results are shown in Table 6. Besides the resulting keyword accuracy, we also display the noise level, the speech level, and the SNR[4] for each of the 6 subsets of the original and modified development and test sets. The results suggest that the Lombard effect has a major impact on ASR performance. By taking it into account in the data simulation instead of assuming a fixed speech level, the SNR and the keyword accuracy in the most noisy test subset increase from -6 dB and 49% accuracy to 5 dB and 82% accuracy, respectively. The SNR and the performance on the less noisy subset also increase to a lesser extent. Overall, the keyword accuracy still increases with the SNR but the absolute accuracy and the rate of increase are quite different. Note that, besides the increase in speech level, other aspects of the Lombard effect are still not taken into account here however their impact on ASR performance is much smaller given suitable feature normalization [51].

5. CONCLUSIONS AND FUTURE WORK

We studied speech distortion conditions in real scenarios for speech processing applications. Information about the acoustic characteristics of different scenarios is quite sparse in the literature. This work attempts to encourage the research community to contribute with their experience to standardizing the dataset process selection, such that the practical value of scientific results is increased. Starting from the identification of the main problems with simulated or real data used for experiments, followed by a summary of the important characteristics of real scenarios to be taken into account, including the properties of reverberation, noise and Lombard effect, we proposed a tentative procedure for quantifying the acoustic properties of each application scenario. This procedure is complementary to the collection of real data. It is useful both for guiding the creation of new real or simulated datasets and for critically analyzing the results reported on existing datasets. We illustrate on a robust ASR task that speech processing performance can drastically chance depending how these characteristics are taken into account in the data.

[3]This level was fixed by recording utterances from a real talker in quiet conditions in the same environment. These utterances are not distributed as part of the publicly available dataset.

[4]Following the convention in [28], the SNR is computed from high-pass filtered versions of the signals with a cutoff frequency of 80 Hz. It is therefore different from the difference between the A-weighed speech and noise levels.

6. REFERENCES

[1] J. M. Baker, L. Deng, J. Glass, S. Khudanpur, C.-H. Lee, N. Morgan, and D. O'Shaughnessy, "Research developments and directions in speech recognition and understanding, part 1," *IEEE Signal Processing Magazine*, vol. 26, no. 3, pp. 75–80, May 2009.

[2] I. Cohen, J. Benesty, and S. Gannot, Eds., *Speech processing in modern communication: Challenges and perspectives*. Springer, 2010.

[3] K. S. Rao and S. Sarkar, *Robust Speaker Recognition in Noisy Environments*. Springer Science+Business Media, 2014.

[4] J. Li, L. Deng, R. Haeb-Umbach, and Y. Gong, *Robust Automatic Speech Recognition — A Bridge to Practical Applications*. Elsevier, 2015.

[5] B. Lee, M. Hasegawa-Johnson, C. Goudeseune, S. Kamdar, S. Borys, M. Liu, and T. Huang, "AVICAR: audio-visual speech corpus in a car environment," in *Proc. Interspeech*, 2004, pp. 2489–2492.

[6] S. Renals, T. Hain, and H. Bourlard, "Interpretation of multiparty meetings: The AMI and AMIDA projects," in *Proc. HSCMA*, 2008, pp. 115–118.

[7] A. Stupakov, E. Hanusa, D. Vijaywargi, D. Fox, and J. Bilmes, "The design and collection of COSINE, a multi-microphone in situ speech corpus recorded in noisy environments," *Computer Speech and Language*, vol. 26, no. 1, pp. 52–66, 2011.

[8] C. Fox, Y. Liu, E. Zwyssig, and T. Hain, "The Sheffield wargames corpus," in *Proc. Interspeech*, 2013, pp. 1116–1120.

[9] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, "The third 'CHiME' speech separation and recognition challenge: Dataset, task and baselines," in *Proc. ASRU*, 2015, pp. 504–511.

[10] M. P. Cooke, J. R. Hershey, and S. J. Rennie, "Monaural speech separation and recognition challenge," *Comput. Speech Lang.*, vol. 24, no. 1, pp. 1–15, 2010.

[11] H.-G. Hirsch and D. Pearce, "The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in *Proc. ASR2000*, 2000, pp. 181–188.

[12] Y. Lei, L. Burget, L. Ferrer, M. Graciarena, and N. Scheffer, "Towards noise-robust speaker recognition using probabilistic linear discriminant analysis," in *Proc. ICASSP*, 2012, pp. 4253–4256.

[13] I. Peer, B. Rafaely, and Y. Zigel, "Reverberation matching for speaker recognition," in *Proc. ICASSP*, 2008, pp. 4829–4832.

[14] E. Vincent, S. Araki, and P. Bofill, "The 2008 signal separation evaluation campaign: A community-based approach to large-scale evaluation," in *Proc. ICA*, 2009, pp. 734–741.

[15] A. El-Solh, A. A. Cuhadar, and R. A. Goubran, "Evaluation of speech enhancement techniques for speaker identification in noisy environments," in *Proc. ISMW*, 2007, pp. 235–239.

[16] S. O. Sadjadi and J. H. Hansen, "Assessment of single-channel speech enhancement techniques for speaker identification under mismatched conditions," in *Proc. Interspeech*, 2010, pp. 2138–2141.

[17] S. Nakamura, K. Hiyane, F. Asano, T. Nishiura, and T. Yamada, "Acoustical sound database in real environments for sound scene understanding and hands-free speech recognition," in *Proc. LREC*, 2000.

[18] C. Hanilci, T. Kinnunen, R. Saeidi, J. Pohjalainen, P. Alku, F. Ertas, J. Sandberg, and M. Hansson-Sandsten, "Comparing spectrum estimators in speaker verification under additive noise degradation," in *Proc. ICASSP*, 2012, pp. 4769–4772.

[19] D. Martinez, L. Burget, T. Stafylakis, Y. Lei, P. Kenny, and E. Lleida, "Unscented transform for ivector-based noisy speaker recognition," in *Proc. ICASSP*, 2014, pp. 4042–4046.

[20] S. Sarkar and K. S. Rao, "Stochastic feature compensation methods for speaker verification innoisy environments," *Journal of Applied Soft Computing*, vol. 19, pp. 198–214, 2014.

[21] W. B. Kheder, D. Matrouf, J.-F. Bonastre, M. Ajili, and P.-M. Bousquet, "Additive noise compensation in the i-vector space for speaker recognition," in *Proc. ICASSP*, 2015, pp. 4190–4194.

[22] Y. Hu and P. C. Loizou, "Subjective comparison and evaluation of speech enhancement algorithms," *Speech Communication*, vol. 49, no. 7–8, pp. 588–601, 2007.

[23] Y. Lei, M. McLaren, L. Ferrer, and N. Scheffer, "Simplified VTS-based i-vector extraction in noise robust speaker recognition," in *Proc. ICASSP*, 2014, pp. 4065–4069.

[24] W. B. Kheder, D. Matrouf, P.-M. Bousquet, J.-F. Bonastre, and M. Ajili, "Robust speaker recognition using MAP estimation of additive noise in i-vectors space," in *Proc. SLSP*, 2014, pp. 97–107.

[25] H.-G. Hirsch, "Aurora-5 experimental framework for the performance evaluation of speech recognition in case of a hands-free speech input in noisy environments," Niederrhein Univ. of Applied Sciences, Tech. Rep., 2007.

[26] Y. Lei, L. Burget, L. Ferrer, M. Graciarena, and N. Scheffer, "A noise robust i-vector extractor using VTS for speaker recognition," in *Proc. ICASSP*, 2013, pp. 6788–6791.

[27] C. Yu, G. Liu, S. Hahm, and J. Hansen, "Uncertainty propagation in front end factor analysis for noise robust speaker recognition," in *Proc. ICASSP*, 2014, pp. 4017–4021.

[28] E. Vincent, J. Barker, S. Watanabe, J. Le Roux, F. Nesta, and M. Matassoni, "The second 'CHIME' speech separation and recognition challenge: Datasets, tasks and baselines," in *Proc. ICASSP*, 2013, pp. 126–130.

[29] D. Ribas, E. Vincent, and J. R. Calvo, "Full multicondition training for robust i-vector based speaker recognition," in *Proc. Interspeech*, 2015, pp. 1057–1061.

[30] T. D. Rossing, Ed., *Springer Handbook of Acoustics*. Springer, 2007.

[31] E. A. P. Habets, S. Gannot, and I. Cohen, "Late reverberant spectral variance estimation based on a statistical model," *IEEE Signal Processing Letters*, vol. 16, no. 9, pp. 770–773, 2009.

[32] P. A. Naylor and N. D. Gaubitch, Eds., *Speech Dereverberation*. London: Springer-Verlag, 2010.

[33] C. F. Eyring, "Reverberation time in "dead" rooms," *Journal of the Acoustical Society of America*, vol. 1, no. 2A, pp. 217–241, 1930.

[34] M. Gardner, "Factors affecting individual and group levels in verbal communication," *Journal of the AES*, vol. 19, pp. 560–569, 1971.

[35] J. Eaton, N. D. Gaubitch, A. H. Moore, and P. A. Naylor, "The ACE challenge — Corpus description and performance evaluation," in *Proc. WASPAA*, 2015, pp. 1–5.

[36] J. H. Rindel, "Verbal communication and noise in eating establishments," *Applied Acoustics*, vol. 71, pp. 1156–1161, 2010.

[37] ——, "Acoustical capacity as a means of noise control in eating establishments," in *Proc. Joint Baltic-Nordic Acoustics Meeting*, 2012.

[38] T. Nishiura, M. Nakayama, Y. Denda, N. Kitaoka, K. Yamamoto, T. Yamada, S. Tsuge, C. Miyajima, M. Fujimoto, T. Takiguchi, S. Tamura, S. Kuroiwa, K. Takeda, and S. Nakamura, "Evaluation framework for distant-talking speech recognition under reverberant environments: newest part of the CENSREC series," in *Proc. LREC*, 2008, pp. 1828–1834.

[39] M. Jeub, M. Schäfer, and P. Vary, "A binaural room impulse response database for the evaluation of dereverberation algorithms," in *Proc. DSP*, 2009, pp. 1–5.

[40] L. Cristoforetti, M. Ravanelli, M. Omologo, A. Sosi, A. Abad, M. Hagmüller, and P. Maragos, "The DIRHA simulated corpus," in *Proc. LREC*, 2014, pp. 2629–2634.

[41] N. Bertin, E. Camberlein, E. Vincent, R. Lebarbenchon, S. Peillon, É. Lamandé, S. Sivasankaran, F. Bimbot, I. Illina, A. Tom, and S. Fleury, "A French corpus for distant-microphone speech processing in real homes," in *Proc. Interspeech*, 2016.

[42] M. Lincoln, I. McCowan, J. Vepa, and H. K. Maganti, "The multi-channel Wall Street Journal audio visual corpus (MC-WSJ-AV): specification and initial experiments," in *Proc. ASRU*, 2005, pp. 357–362.

[43] H. Knecht, P. Nelson, G. Whitelaw, and L. Feth, "Background noise levels and reverberation times in unoccupied classrooms," *American Journal of Audiology*, vol. 11, pp. 65–71, 2002.

[44] C. Crandell and F. Bess, "Speech recognition of children in a typical classroom setting," *American Speech-Language and Hearing Association*, vol. 29, pp. 907–939, 1987.

[45] T. Finitzo-Hieber, "Classroom acoustics," in *Auditory disorders in school children*, R. Roeser and M. Downs, Eds. New York: Theime-Stratton, 1988, pp. 221–233.

[46] K. S. Pearsons, R. L. Bennett, and S. Fidell, "Speech levels in various noise environments," Office of Research and Development, Environmental Protection Agency, Washington, USA, Tech. Rep. EPA-600/ 1-77-q2-5, 1977.

[47] E. van Heusden, R. Plomp, and L. C. W. Pols, "Effect of ambient noise on the vocal output and the preferred listening level of conversational speech," *Journal of Applied Acoustics*, vol. 12, pp. 31–43, 1979.

[48] J. Thiemann, N. Ito, and E. Vincent, "The Diverse Environments Multi-channel Acoustic Noise Database (DEMAND): A database of multichannel environmental noise recordings," in *Proc. 21st Int. Congress on Acoustics*, 2013, pp. 1–6.

[49] H. Lazarus, "Prediction of verbal communication in noise — A review: Part 1," *Applied Acoustics*, vol. 19, pp. 439–464, 1986.

[50] J. H. L. Hansen, A. Sangwan, and W. Kim, "Speech under stress and Lombard effect: Impact and solutions for forensic speaker recognition," in *Forensic Speaker Recognition: Law Enforcement and Counter-Terrorism*, A. Neustein and H. A. Patil, Eds. New York: Springer, 2012, pp. 103–123.

[51] J. H. L. Hansen, "Analysis and compensation of speech under stress and noise for environmental robustness in speech recognition," *Speech Communication*, vol. 20, pp. 151–173, 1996.

[52] J. H. L. Hansen and V. Varadarajan, "Analysis and compensation of Lombard speech across noise type and levels with application to in-set/out-of-set speaker recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 2, pp. 366–378, 2009.

[53] "ISO 9921. Ergonomics — Assessment of speech communication," Geneva, 2003.

[54] J. Garofalo, D. Graff, D. Paul, and D. Pallett, "CSR-I (WSJ0) Complete," Philadelphia, 2007.

[55] M. P. Cooke, J. Barker, S. P. Cunningham, and X. Shao, "An audio-visual corpus for speech perception and automatic speech recognition," *Journal of the Acoustical Society of America*, vol. 120, pp. 2421–2424, 2006.

[56] S. J. Young, G. Evermann, M. J. F. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. C. Woodland, *The HTK Book, version 3.4*, University of Cambridge, 2006.