A speech database for the study of Lombard effect in Brazilian Portuguese

David Daniel e Silva
DCC/UDESC
Joinville/SC, Brazil
Email: david@das.ufsc.br

Carlos Alberto Ynoguti
Instituto Nacional de Telecomunicações
Santa Rita do Sapucaí, MG, Brazil
Email: ynoguti@inatel.br

Marcelo Ricardo Stemmer
PPGEAS/DAS/UFSC
Florianópolis/SC, Brazil
Email: marcelo@das.ufsc.br

***Abstract*—The Lombard effect is related to the change in the way people speak in noisy environments in order to compensate the degradation of the audio signal. There are significant changes in the duration of some types of phones and in the formant positions that can vary from a language to another. The construction of a speech database that isolates the Lombard effect is a challenging task because its necessary to record only the speech signal, without the environment noise. In this paper a methodology to construct such databases is described and a database for the Brazilian Portuguese was produced using this procedure.**

***Index Terms*—Automatic speech recognition, Lombard effect, speech database.**

I. INTRODUCTION

Under noisy conditions people use to talk in a different way, trying to compensate the degradation of the audio signal. This is called Lombard effect, and is usually defined as the variation of speech articulation produced by the speaker in an attempt to improve the intelligibility [1].

There are studies to characterize the Lombard effect in many languages but, until the moment, the authors don't have notice of any initiative of this kind for the Brazilian Portuguese.

Several studies show that for automatic speech recognition in noisy environments, the word error rate depends not only on the noise level but also on the intensity of the Lombard effect [1]. Additionally, the great increase in the use of mobile devices for communication make this effect more and more present in actual situations. Thus it is important to have a good understanding of this effect in order to investigate ways to compensate the word error rate degradation of such systems.

The present work focuses on a methodology to construct a database with speech produced when the speaker is trying to communicate in noisy environments, and therefore experiencing the Lombard effect. The challenge here is that it is necessary to record only the speech but the recording environment must be contaminated with noise.

In this direction much research has been made to improve the communication due to Lombard effect. In [2] for example, a low-vocabulary speech recognition algorithm that provides robust performance in noisy environments with particular emphasis on characteristics due to the Lombard effect is described. The database used was the Speech Under Simulated and Actual Stress (SUSAS) [3], and can be obtained at the Linguistics Data Consortium (LDC) [4].

Another initiative is reported in [5]. In this work, it is argued the study of the Lombard reflex requires more realistic databases, and then utterances representing people speaking in real-world conditions need to be recorded. To accomplish this goal, 5 male and 5 female subjects were recorded in 8 different scenarios: when reading a list of 50 phrases (comprised of first and/or last names) in quiet and with 3 different types of noise, differing mainly by their spectral tilt, and when talking to a voice dialing system (which was trained with the list of 50 phrases in quiet) in quiet and in the three noise conditions. The vocabulary was chosen to include most of the American English phonemes. During the experiments involving recognition, the speakers marked a score sheet indicating if the recognizer correctly recognized the first, second or third candidate. For the 8 different scenarios the vocabulary was randomized and 5 phrases were added at the beginning of the list for the subject to adapt to the experiment. The database was manually labeled at the phoneme level and digitalized at 16 kHz sampling rate.

Thus, to contribute with the studies about Lombard effect, this work presents a methodology to produce a Lombard speech database. Furthermore, a database comprising 44 adult speakers was produced with the proposed methodology.

The paper is organized as follows: in Section II a brief description of Lombard effect is given. Section III outlines the methodology used to build the database. The experimental setup is described in Section IV, whereas in Section V the speakers and utterance types are presented. In section VI the noise types and levels are shown, and in Section VII the recording procedure is detailed. Finally, section VIII presents the conclusions and future work.

II. THE LOMBARD EFFECT

The effect was discovered in 1909 by ´Etienne Lombard, a French otolaryngologist. As stated above, the Lombard effect or Lombard reflex is the involuntary tendency of speakers to increase the intensity of their voice when speaking in noisy environment to enhance its audibility [6].

Changes between normal and Lombard speech include [7][8]:

- increase in phonetic fundamental frequencies
- shift in energy from low frequency bands to middle or high bands

- increase in sound intensity
- increase in vowel duration
- spectral tilting
- shift in formant center frequencies for F1 (mainly) and F2

The auditory feedback is also pointed as an important factor or the Lombard speech, because its primary goal is to improve the intelligibility in noisy conditions [9]. Also, it was found that the intelligibility increases up to a certain level and then drops dramatically when the speaker starts to shout [1].

Additionally, a study made by Hansen showed that the Lombard effect varies with the type and intensity of the noise [1]. This work used the SUSAS database too.

III. Methodology

For the study of the Lombard effect, one is interested in pronunciation change induced by the environmental noise [1][10]. However, the isolation of the Lombard effect from the background noise is not a trivial task, because the speaker must talk in a noisy environment but the recordings must have only the speech signal, without the noise.

The first approach that was considered for this task was to make the speakers listen to the noises through earphones. With this choice, it would be easy to capture only the speech signal, without the noise. But this approach has two drawbacks: a) it is difficult to measure the noise power at the output of the earphones and b) the perception of the noise would be more realistic with the use of loudspeakers than through earphones.

As an alternative, we could use a throat-microphone, which is less sensitive to the environment noise. Evaluation of preliminary recordings, with 3 male and 3 female adult speakers, showed that the environmental noise was greatly reduced with this setup.

Considering all these aspects, the last approach was chosen, with the reasoning that it is preferable to tolerate a small amount of noise and maintain the “real” influence of the noise on the speakers.

IV. Experimental Setup

As mentioned above, a throat-microphone was used for the recordings. The model used was a Motorola Talkabout T5025 [11], which is shown in Figure 1.

Fig. 1. Microphone used for recordings.

The recordings were performed in a semi-anechoic room, with the walls and roof covered with a high-density foam. This room has its acoustic performance assessed according to the international standard ISO 3745 [12] procedure. As the SPL (Sound Pressure Level) can vary according to the average absorption coefficient of the walls [13], the knowledge of the acoustic properties of the recording room is important when making power measurements of the noises. Figure 2 shows details of this room.

Fig. 2. Semi-anechoic room used for recordings.

It is a known fact that the perceived sound level does not change significantly for distances less than 1 m from the sound source [14]. Also, the surround effect is more realistic if the audio is played through more than one speaker, positioned behind the listener [15]. With these facts in mind, the noises were played through 3 loudspeakers positioned 0.5 m behind the speakers. Figure 3 shows the loudspeakers positioning details.

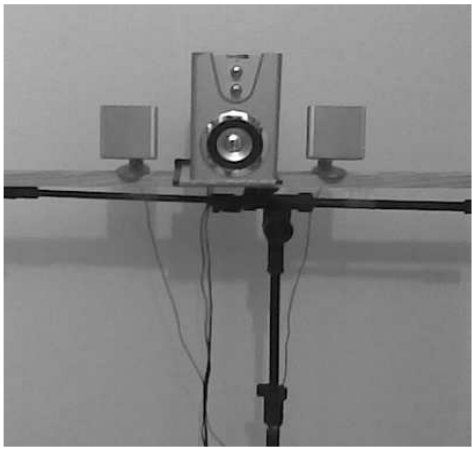

Fig. 3. Loudspeakers positioning.

The sound pressure was measured near the ears of the speaker with a digital sound level meter. The characteristics of this device are presented below:

- scales: 30 to 130 db (3 scales)
- resolution: 0,1 dB
- exactness: 1,4 dB
- dynamic reach: 50 dB
- frequency: 31.5 kHz to 8 kHz
- norm IEC-61672 type 2 and ANSI S1.4 type 2
- detachable capacitive microphone
- fast and slow reply
- curve of reply A and C

In Figure 4 we have an image of the sound level meter used in this work.

Fig. 4. Digital sound level meter used to adjust the noise volume.

V. Speakers and Utterance Types

For the recordings, 44 adult speakers (22 male and 22 female) were selected.

Each speaker pronounced 20 phonetically balanced sentences for the Brazilian Portuguese. These sentences correspond to the lists 9 and 12 proposed in [16] (each list has 10 sentences). One interesting feature of these lists is that every set of 10 sentences has at least one occurrence of all phones for the Brazilian Portuguese. Thus it is possible to cover the main variabilities of the language with a small amount of recordings, which is a crucial detail when the speakers must be submitted to high levels of noise during the recording sessions. The sentences used for this database are shown in Table I.

TABLE I.
Brazilian Portuguese Sentences (From [16]).

1	Os maiores picos da Terra ficam debaixo d'água.
2	A inauguração da vila é quarta-feira.
3	Só vota quem tiver o título de eleitor.
4	É fundamental buscar a razão da existência.
5	A temperatura só é boa mais cedo.
6	Em muitas regiões a população está diminuindo.
7	Nunca se pode ficar em cima do muro.
8	Pra quem vê de fora o panorama é desolador.
9	É bom te ver colhendo flores.
10	Eu me banho no lago ao amanhecer.
11	As crianças conheceram o filhote de ema.
12	A bolsa de valores ficou em baixa.
13	O Congresso volta atrás em sua palavra.
14	A médica receitou que eles mudassem de clima.
15	Não é permitido fumar no interior do ônibus.
16	A apresentação foi cancelada por causa do som.
17	Uma garota foi presa ontem à noite.
18	O prato do dia é couve com atum.
19	Eu viajarei ao Canadá amanhã.
20	A balsa é o meio de transporte daqui.

Additionally, 20 sequences of isolated and connected digits from the Aurora 1 database [17] were also pronounced by each speaker. These utterances can be useful to study the Lombard effect among non-native english speakers. They are shown in Table II.

TABLE II.
Sequences of digits (From [17])

1	oh
2	zero
3	one
4	two
5	three
6	four
7	five
8	six
9	seven
10	eight
11	nine
12	eight six zero one one six two
13	oh oh two one six four one
14	four three oh six five seven one
15	nine eight zero seven four three seven
16	seven zero four six nine eight nine
17	zero four seven four five six three
18	nine nine eight nine two seven six
19	two four seven three four eight zero
20	five zero five five three nine zero

VI. Noise Types and Levels

Four noises were used in this work: metal-cutting, tunnel-front, tunnel-inside and crowd-children.

The metal-cutting noise has been recorded while a dry multi-cutting saw was processing a 3/4" thickness stainless steel block. The microphone was placed at a distance of approximately 3 m from the machine, inside a masonry shed with around 10 m width, 30 m length and 6 m height.

The tunnel-front noise was obtained in a highway, around 15 m away from the tunnel's entrance and 2 m away from the track.

The tunnel-inside noise was obtained at approximately the half of the tunnel's length.

The crowd-children noise was obtained at a school with a group of scholars between 5 and 12 years old at the classroom arrival time.

In all recordings, the microphone used was a flexible omnidirectional with adjustable headset, specially developed for multimedia applications, speech recognition, language laboratories, etc. The microphone specifications/characteristics are: impedance: 32 ohms; sensitivity: 96 dB +/- 4 dB; frequency range: 20 Hz to 20 kHz; polar pattern: omnidirectional. It was placed at 2 m height in all recording environments.

All the noises were recorded with a notebook using professional audio/speech editing software, with a sampling rate of 16 kHz and 16 bit quantization in one channel. The sampling rate was converted for 8 kHz, the same used in Aurora-1 database.

VII. Recording Procedure

Each speaker pronounced the 20 sentences and the 20 sequences of connected digits in each of the noise situations below:

- without noise
- metal-cutting noise with 70 dB, 80 dB and 90 dB
- tunnel-front noise with 70 dB, 80 dB and 90 dB
- tunnel-inside noise with 70 dB, 80 dB and 90 dB
- crowd-children noise with 70 dB, 80 dB and 90 dB

Thus the speakers pronounced each utterance 13 times, leading to a total of 40x13=520 utterances. With 44 speakers, the whole database consists of 44x520=22880 utterances.

The procedure adopted for the recording sessions was: with the speaker sitting on a chair, wearing the throat-microphone the loudspeakers were positioned at an appropriate height. In the sequence, the noise was played through the loudspeakers and its level was adjusted according to the desired level using the digital sound level meter. Finally, the speaker read all the 40 utterances without interruptions. The individual sentences were separated after the recording session. After this first sequence, the noise level and/or type was changed and the procedure was repeated until all the noises were played in the 3 levels (70 dB, 80 dB and 90 dB).

This procedure was chosen in order to minimize the time that the speaker would be submitted to the noise, but even with this cautions a typical recording session lasts about 50 min. This fact prevented many people to participate in this project even with the promise of a delicious chocolate bar at the end of the session.

VIII. Conclusions and Future Work

In this work a methodology for the construction of databases for the study of Lombard effect on speech was described. Also, a database produced under this methodology was presented.

This database consists of recordings of 44 adult speakers (22 male and 22 female), collected under four different types of noise (metal-cutting, tunnel-front, tunnel-inside and crowd-children) and 3 levels (70 dB, 80 dB and 90 dB).

The recordings were performed in a semi-anechoic room with a throat-microphone and with the noises being played through loudspeakers to improve the sensation of real noises for the speakers.

20 phonetically balanced sentences for the Brazilian Portuguese and 20 sequences of isolated and connected digits in English were used. The first set was chosen to make studies about the Lombard effect for the Brazilian Portuguese Language, and the second set will be used to study the Lombard effect among non native English speakers.

For the future, we intend to study the influence of the Lombard effect in automatic speech recognition systems for the Brazilian Portuguese.

Acknowledgment

The authors would like to thank UDESC for the financial support, UFSC/PPGEAS for the physical structure and INATEL for orientation and technical support.

References

[1] Hansen, J. H. L.; Varadarajan, V. 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, February 2009.

[2] Chi, S.M., Oh, Y.-H., Lombard Effect Compensation and Noise Suppression for Noisy Lombard Speech Recognition, Proceedings of the Fourth International Conference on Spoken Language Processing. Vol 4., Philadelphia, PA. October 1996. pp. 2013-2016.

[3] Hansen, J. H. L.; Bou-Ghazale, S. Getting started with Susas: A speech under simulated and actual stress database, Proceedings of the Eurospeech' 97, Rhodes, Greece, Sep. 1997, vol. 4, p. 1743-1746. [4]

[4] http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC99S78 (Access: February 2010).

[5] Junqua, J. C., Fincke, S. and Field, K., The Lombard Effect: A Reflex to Better Communicate With Others in Noise, Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing,1999. Vol. 04. Phoenix, Arizona. March 1999. pp. 2083-2086.0202

[6] Lane, H. L.; Tranel, B. The Lombard sign and the role of hearing in speech, Journal of Speech and Hearing Research, vol. 14, pp. 677-709, 1971.

[7] Junqua J. C. The Lombard reflex and its role on human listeners and automatic speech recognizers. Journal of the Acoustical Society of America 93 (1), January 1993. pp. 510-524.

[8] Summers W. V., Pisoni D. B., Bernacki R. H., Pedlow R. I., Stokes M. A. Effects of noise on speech production: acoustic and perceptual analyses. Journal of the Acoustical Society of America 84 (3).(September 1988). pp. 917-928.

[9] Hansen, J. H. L. Analysis and Compensation of Stressed and Noisy Speech with Application to Robust Automatic Recognition, Ph.D. dissertation, School of Electrical Engineering, Georgia Institute of Technology, Atlanta, 1988.

[10] Stanton, B. J.; Jamieson, L. H.; Allen, G. D. Acoustic-phonetic analysis of loud and Lombard speech in simulated cockpit conditions, Proceedings of the International Conference on Acoustics, Speech and Signal Processing. 1988, pp. 331-334.

[11] www.motorola.com (Access: October 20th 2009).

[12] ISO 3745, Acoustics C Determination of sound power levels of noise sources C Precision methods for anechoic and semi-anechoic rooms, International Organization for Standardization, 1977.

[13] Ballagh, Qualification tests in a semi-anechoic room, Acoustical, 1984, Vol. 54, Research Note, pp. 296-299.

[14] Walshaw, A. C. Mechanical Vibrations With Applications. Ellis Horwood Limited, 1984.

[15] Li, Z.; Duraiswami, R.; Davis, L. S. Recording and Reproducing High Order Surround Auditory Scenes for Mixed and Augmented Reality, Computer Society, Proceedings of the Third IEEE and ACM International Symposium on Mixed and Augmented Reality (Ismar 2004).

[16] Alcaim, A.; Solewicz, J. A.; Moraes, J. A. Frequência de ocorrência dos fones e lista de frases foneticamente balanceadas no português falado no Rio de Janeiro. Revista da Sociedade Brasileira de Telecomunicações. P. 23-41. Dezembro, 1992.

[17] Hirsch, H.G.; Pearce, D. The Aurora Experimental Framework for the Performance Evaluation of Speech Recognition Systems under Noisy Conditions, ISCA ITRW ASR2000, Paris, France, September 2000.