THE LOMBARD EFFECT: A REFLEX TO BETTER COMMUNICATE WITH OTHERS IN NOISE

Jean-Claude Junqua, Steven Fincke and Ken Field

Panasonic Technologies, Inc. / Speech Technology Laboratory, 3888 State Street, Suite #202
Santa Barbara, California, 93105, U.S.A.
Tel. (805) 687-0110; fax: (805) 687-2625; email: jcj@research.panasonic.com

ABSTRACT

To study the Lombard reflex, more realistic databases representing real-world conditions need to be recorded and analyzed. In this paper we 1) summarize a procedure to record Lombard data which provides a good approximation of realistic conditions, 2) present an analysis per class of sounds for duration and energy of words recorded while subjects are listening to noise through open-ear headphones a) when speakers are in communication with a recognition device and b) when reading a list, and 3) report on the influence of speaking style on speaker-dependent and speaker-independent experiments. This paper extends a previous study aimed at analyzing the influence of the communication factor on the Lombard reflex. We also show evidence that it is difficult to separate the speaker from the environment stressor (in this case the noise) when studying the Lombard reflex. The main conclusion of our pilot study is that the communication factor should not be neglected because it strongly influences the Lombard reflex.

1. INTRODUCTION

In the presence of noise, speech is masked, and its production is modified by what is known as the Lombard reflex [1, 2, 3]. At the speech recognition level, this reflex has been mostly neglected due to the difficulties in characterizing this reflex and to its speaker-dependency. A number of acoustic-phonetic studies have shown that some parameters such as pitch, formant frequencies, duration and energy distribution are affected by the non-linear changes in speech production when speech is produced in noise as compared to speech production in quiet conditions. However, it is still difficult to incorporate this partial understanding in automatic speech recognizers. So far, the most popular methods to deal with this reflex in automatic speech recognition have been multi-style training, simulated Lombard token generation and feature compensation methods [4]. An alternative to the development of robust algorithms is helping the user to regulate his/her voice by increasing the speaker sidetone or providing visual feedback.

In this paper, we are exploring how the communication factor influences speech produced in noise. It has been implicitly assumed that the Lombard reflex is a physiological effect. However, it seems that in the real world, the magnitude of the response of the speakers is governed by the desire to achieve successful intelligible communication [5, 6]. Most current databases do not place emphasis on communication, although this seems to be an important factor to consider; as noted in [7], "The speaker does not change his voice level to communicate better with himself, but rather with others." In [8] we presented a study which showed that the Lombard reflex varies depending on whether the subject is in communication or reading a list. After reviewing the recording procedure, this paper follows up on this previous study by presenting detailed results at the phoneme-class level, while emphasizing the inter-speaker variability in magnitude of Lombard response which may be due to individuals' differences in coping with environmental noise.

2. A RECORDING PROCEDURE PROVIDING A GOOD APPROXIMATION OF REAL-WORLD CONDITIONS

As shown in Figure 1., to assess the influence of the communication factor on the Lombard reflex and automatic speech recognition, we recorded a database using a telephone containing a prototype of a speaker-dependent automatic speech recognizer for voice dialing.

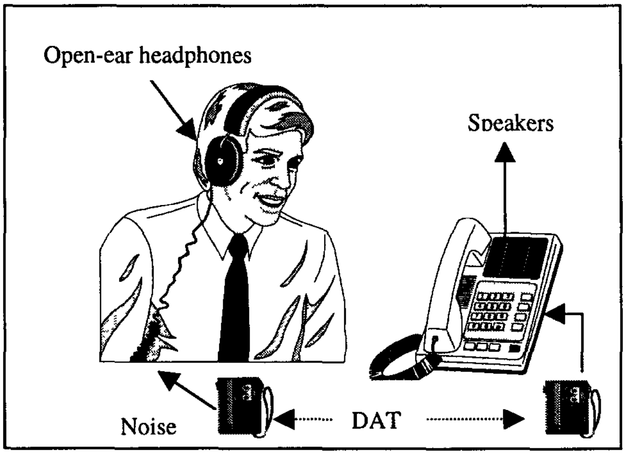

Figure 1: The recording setup.

In all the experiments, the user's speech was recorded on digital audio tape (Panasonic SD-DA10). The subjects spoke into the phone using the handset. The audio output from the speakerphone was channeled to multi-media speakers during recognition training and testing. Subjects were allowed to adjust the volume of the speakers. The volume was usually higher when there was noise. For all the recording conditions, *subjects were wearing open-ear headphones* (Sennheiser HD 580), which, in the experiments involving noise, were used to inject noise to the subjects at 85 dB SPL. By using open-ear headphones the subjects were able to hear the audio output from the multimedia speakers without any sound attenuation.

3. THE DATABASE

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 (see [8] for more details), and when talking to the 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 subjects marked a score sheet indicating if the recognizer was correct with its 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. In the following sections the three noises will be referred with the names pinklvl, pinktilt and speech, respectively, in order of increasing spectral tilt.

4. A PHONEME CLASS-BASED ANALYSIS OF THE RECORDED DATA

We analyzed the data recorded per class of sound for two parameters particularly important for speech recognition such as duration and spectral power. Figure 2. shows the average duration per class of sounds for male and female speakers and the 8 different scenarios recorded. It can be seen that, for almost all the conditions and classes of sounds (except for glides and plosives in the case of female speakers), the average duration decreases when subjects are in communication with the recognition device as compared to when reading a list.

For the vowels, glides, liquids and nasals, durations were greater for speech production in noise than in quiet. This result is valid when subjects are reading a list or are in communication with the recognition device. For affricates, fricatives and plosives, the data is less conclusive. When noise spectral tilt increases, vowel duration tends to decrease. This is also generally true for liquids, fricatives and affricates. These duration results correlate relatively well with the signal-to-noise ratios (SNRs) of the data in the different conditions (see [8]).

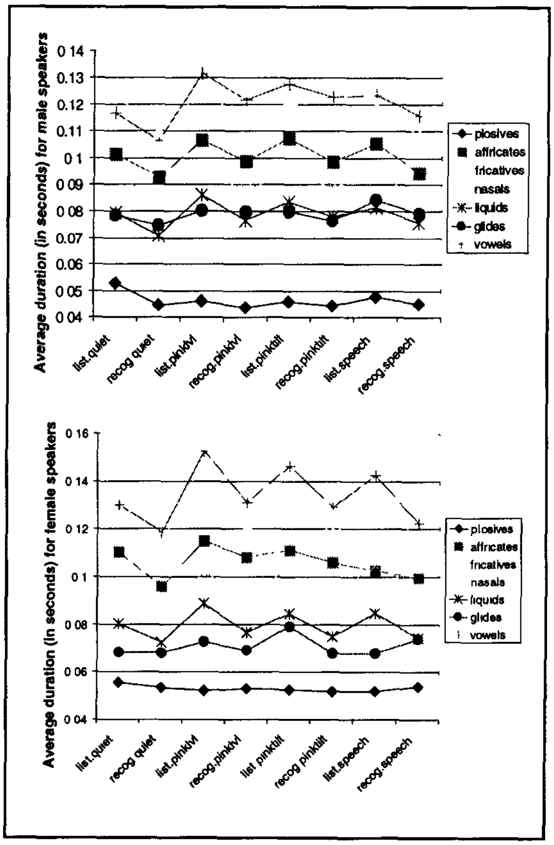

Figure 2. Average duration per class of sounds according to the various recording conditions.

For the energy parameter we divided the frequency bandwidth in three bands: low-band (0-999 Hz), mid-band (1000-3499 Hz) and high-band (3500-8000Hz). Utilizing the manual phoneme-level segmentation, the power spectrum was averaged over the duration of the phoneme. For the conditions when the subjects were in communication with the recognition device, Figure 3. shows the relative power spectrum averaged over all speakers for the three frequency bands. If we take the condition when the subjects were talking to the recognition device in quiet as a reference, it seems that when subjects are producing speech in noise the greatest increase in energy is in the higher frequencies. As already mentioned in previous studies (e.g. [2]) the energy center of gravity is increased when speech is produced in noise as compared to when speech is produced in quiet. When subjects produced speech in presence of noise having a spectral shape similar to the speech spectrum, the increase of energy is less than for the other noises. This is especially true for the high frequency-band. This supports the results of [9] which showed that the energy tends to increase in the frequency bands where the intensity of the noise is relatively high.

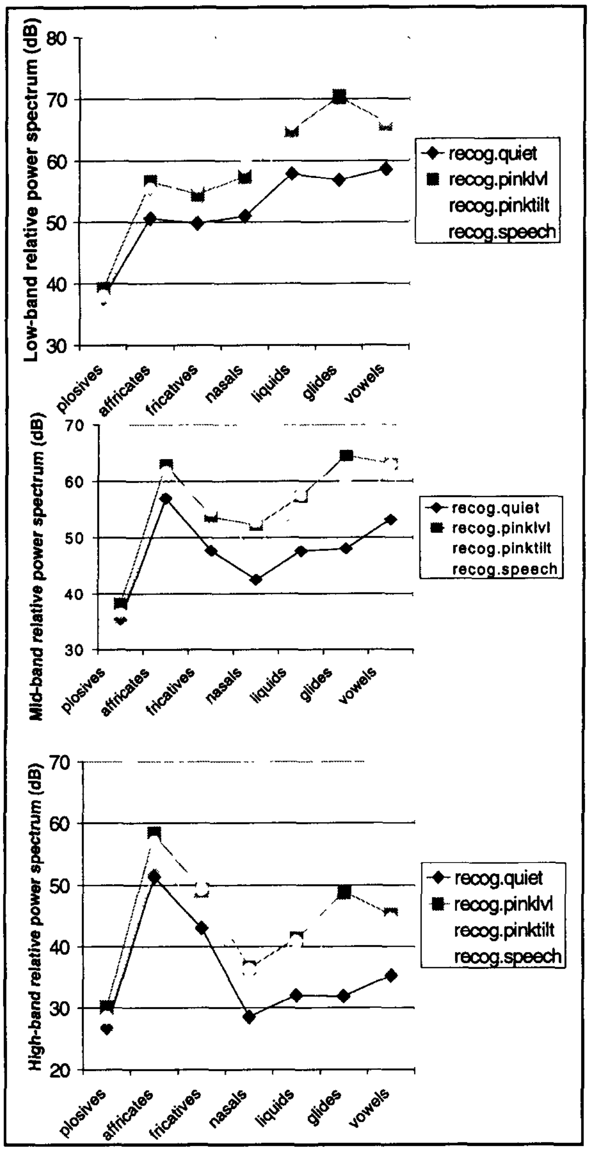

Figure 3. Average over all the speakers of low, mid and high-band power spectrum for data recorded when subjects were in communication with the recognition device.

5. SPEAKER-DEPENDENT AND SPEAKER-INDEPENDENT EXPERIMENTS

In [8] we presented the results of on-line recognition experiments with the recognition device along with off-line speaker-independent recognition performance on the data recorded with different beam thresholds. In this section, we show speaker-dependent and speaker-independent recognition experiments on the recorded database. For the speaker-dependent experiments, the data used to train the phone was used to build the speaker-dependent models while speaker-independent experiments used a flexible vocabulary recognizer with full search and one transcription per word. Both types of experiments were performed on a vocabulary size of 50 words with a total of 500 trials. Figure 4. shows that communicating with the recognition device improves recognition performance as compared to when subjects are reading a list. Moreover, recognition accuracy tends to increase in tandem with the noise spectral tilt. Compared to the results presented in [8], speaker-dependent results are much higher and speaker-independent results presented in [8] did not show a clear improvement of the recognition accuracy when noise spectral tilt increased. The fact that higher performance is obtained in the speaker-dependent tests is due to an updated version of the recognizer used for the results in Figure 4. and some differences between on-line and off-line tests. In particular off-line tests did not used ulaw quantization and codec filtering. Furthermore, a number of errors during on-line tests were due to the endpoint detector which failed to detect speech for low level signals. As for the speaker-independent experiments, differences may be due to the interaction between a number of parameters such as the vocabulary size (which was 662 in our previous experiments) and the transcription quality. However, these results support the claim that the Lombard reflex depends on both 1) the noise type and 2) the fact that a subject is in communication with a recognition device or not.

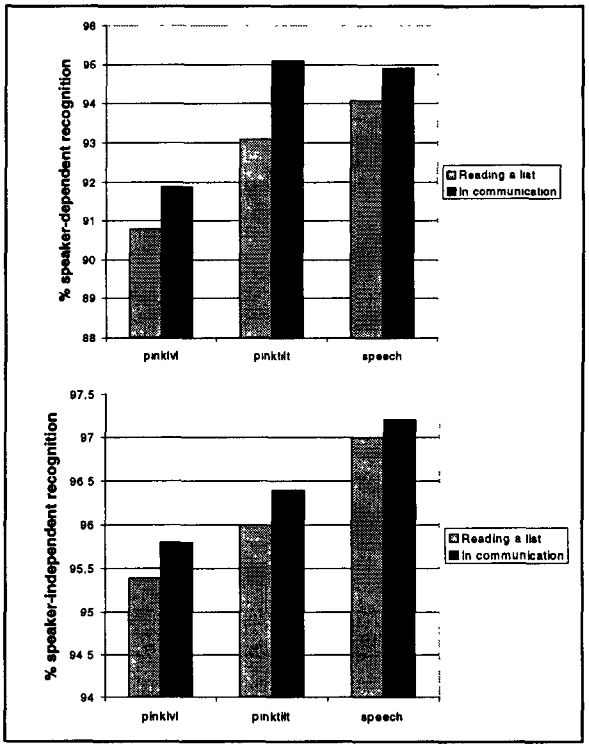

Figure 4. Average off-line recognition accuracy over all the speakers for speaker-dependent and speaker-independent experiments.

6. ARE SPEAKERS COPING DIFFERENTLY WITH NOISE?

In [2] it was suggested that Lombard speech variability may depend on how normal speech is produced. Another hypothesis is that each speaker has his/her own way of coping with noise. The effect of noise on people can be somewhat attenuated by adoption of an appropriate strategy which may be inherent to the person him/herself. This implies that people are, within some limits, able to modify the effects of noise on their speech production. Figure 5. shows how the SNR of the 10 recorded speakers varies across the conditions. It can be seen that for some speakers the standard deviation is quite small (e.g. m.vm and f.jc) and for some others it is quite large (e.g. m.jw and f.fc). This suggests that noise affects speakers differently. Noise/stress can be considered as an adaptation mechanism (physiological) which contributes to a stabilization of our main functions [10]. The idea behind this definition is that stress corresponds to an increase of energy in response to an aggression. Depending of how the aggression is perceived, the adaptation mechanism varies. This is a possible interpretation of the differences between how different speakers are coping with noise.

Due to the ability of the speakers to adapt to their environment (noisy or not) it would be interesting to measure how, in our experiments, the presentation order of the conditions tested affects our results. A random order was selected for each speaker. However, when the total number of speaker is small this effect should not be neglected.

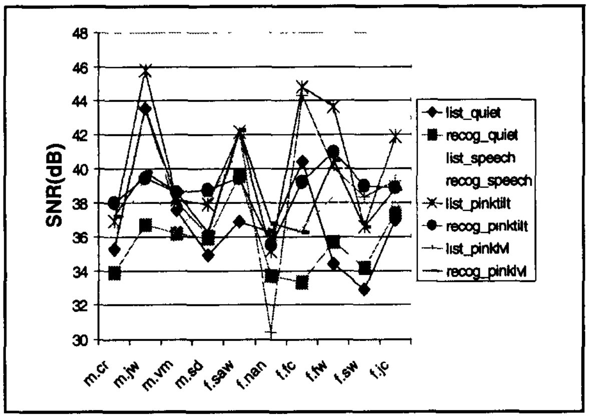

Figure 5. Average SNR over the 50 words recorded for individual speakers across the 8 recorded conditions.

7. CONCLUSIONS

In this paper, we reviewed a procedure providing a good approximation of realistic conditions for recording Lombard data. Instead of recording the analog signal on a DAT tape, this procedure could be improved by dumping the digital signal directly onto a computer disk. This would eliminate the effect of quantization, filtering or audio amplification between on-line and off-line tests. We conducted a phoneme class-based analysis and reported on off-line speaker-dependent and speaker-independent experiments.

The main conclusions of this work lies in the dependence of the Lombard reflex on the speaker's purpose which can be to communicate with others or not (e.g. when reading a list). While the data presented shows that the type of noise is another degree of freedom to take into account in characterizing the Lombard reflex, more data needs to be recorded and analyzed to draw more detailed conclusions.

At the recognition level, depending on the task to be performed, the degradation in recognition accuracy due to the Lombard reflex can be significant or not. The user interface and the influence of the task on the speaker itself contribute to the Lombard reflex. This effect should not be neglected.

8. ACKNOWLEDGEMENTS

The authors would like to acknowledge Man Moham Sondhi from Bell Laboratories, Lucent Technologies for suggesting the use of the open-ear headphones Sennheiser HD 580 to record Lombard speech. They also would like to thank Dr. Philippe Morin for running the speaker-dependent experiments.

9. REFERENCES

1. Lombard E. " Le Signe de l'Elévation de la Voix", Ann. Maladies Oreille, Larynx, Nez, Pharynx, Vol. 37, 1911, pp. 101-119.
2. Junqua J-C. "The Lombard Reflex and its Role on Human Listeners and Automatic Speech Recognizers", J. Acoust. Soc. Am., 1993, Vol. 1, pp. 510-524
3. Junqua J-C. "The Influence of Acoustics on Speech Production: A Noise-Induced Stress Phenomenon Known as the Lombard Reflex", Speech Communication, 1996, Vol. 20, pp. 13-22
4. Hansen J.H.L. "Analysis and Compensation of Speech Under Stress and Noise for Environmental Robustness in Speech Recognition", ESCA/NATO Workshop on Speech Under Stress, September 1995, pp. 91-98.
5. Halphen E. "Des Lésions Traumatiques de l'Oreille Interne", Ph.D. Thesis, Faculté de Médecine, Paris, 1910
6. Egan J.J. " Psychoacoustics of the Lombard Voice Reflex", Ph.D. Thesis, Western Reserve University, 1967
7. Lane H. and Tranel, "The Lombard Sign and the Role of Hearing in Speech" J. Speech and Hearing Research, 1971, Vol. 14, pp. 677-709.
8. Junqua J-C., "Influence of the Speaking Style and the Noise Spectral Tilt on the Lombard Reflex and Automatic Speech Recognition", to be published in ICSLP-98.
9. Mokbel C. " Reconnaissance de la Parole dans le Bruit: Bruitage/Débruitage", Ph.D. Thesis, 1992, Ecole Nationale Supérieure des Télécommunications.
10. Olivier J-F. "Le Stress Moteur de la Vie", Editions Encre, 1991.