INFLUENCE OF THE SPEAKING STYLE AND THE NOISE SPECTRAL TILT ON THE LOMBARD REFLEX AND AUTOMATIC SPEECH RECOGNITION

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) propose a procedure to record Lombard data which provides a good approximation of realistic conditions and 2) present a comparison between two sets of experiments where subjects are in communication with a device while listening to noise through open-ear headphones and where subjects are reading a list. By studying acoustic correlates of the Lombard reflex and performing off-line speaker-independent recognition experiments it is shown that *the communication factor affects the Lombard reflex.*

We also show evidence that *several types of noise differing mainly by their spectral tilt induce different acoustic changes*. This result reinforces the notion that it is difficult to separate the speaker from the environment stressor (in this case the noise) when studying the Lombard reflex.

1. INTRODUCTION

When speech is produced in noise there is a modification of speech production leading to the Lombard reflex [1, 2, 3, 4]. To elicit this reflex, several databases have been recorded while the speakers listened to noise through headphones. Using such a method, a great variability in the increase of vocal effort was observed. This may come from the way databases are recorded. Generally, they implicitly assume 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 obtain intelligible communication [5, 6]. Even if the speakers were asked to speak in such a way as to be intelligible, there was no premium on intelligibility because the speakers were reading a list. Thus, 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." This premium on intelligibility is not universally accepted, however; another hypothesis is that increases in vocal intensity in noise are mainly mediated by an automatic regulating device [8, 9]. Current databases emphasize more an increase in vocal effort due to masking noise than a modification of speech production to be more intelligible by others. In this paper, we contrast the communication speaking style with the reading speaking style and study 1) how several acoustic correlates of the Lombard data recorded in the different conditions vary and 2) how automatic speech recognition (ASR) is influenced by these variations.

Another interesting question lies in the dependence of the Lombard reflex on the type of environment or type of noise. It is now well known that if the noise is stationary an adaptation of the speaker over time to the noisy environment is likely to occur. Several studies hinted that the frequency distribution of the noise affects the Lombard reflex and shapes the acoustic changes observed in the speech signal. In [2], it was found that the type of noise influences vowel duration. In [10], it was reported an increase of the speech energy in the frequency bands where the noise energy is most important. These results suggest that the Lombard reflex depends on the noise frequency distribution. There is also evidence that the noise affects a speaker not only physiologically but also psychologically [11]. Noise will induce different acoustic changes depending on how speakers can cope with it. In fact, when studying the Lombard reflex, it may be difficult to separate the speaker from the environment stressor (in this case the noise). In the current study we evaluate the effect of the noise spectral tilt on several acoustic parameters and automatic speech recognition.

In the following sections we present 1) the recording procedure and a preliminary analysis of the data recorded, 2) shows how recognition accuracy of a speaker-independent speech recognizer is influenced by the Lombard speech variability and 3) briefly discuss our results. The Lombard reflex is very difficult to characterize and furthermore it is very speaker-dependent (e.g. [3]). Consequently, in this paper we will focus on the differences in trend between speaking styles and how recognition accuracy is affected more than on detailed differences which often vary from speaker to speaker.

2. RECORDING OF A LOMBARD DATABASE WHICH PLACES AN EMPHASIS ON COMMUNICATION

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. In all the experiments, the user's speech was simultaneously recorded on a digital audio tape (Panasonic SD-

DA10). 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. 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. 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 Figure 1), 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 to accustom the subject to the experiment. The database was manually labeled at the phoneme level.

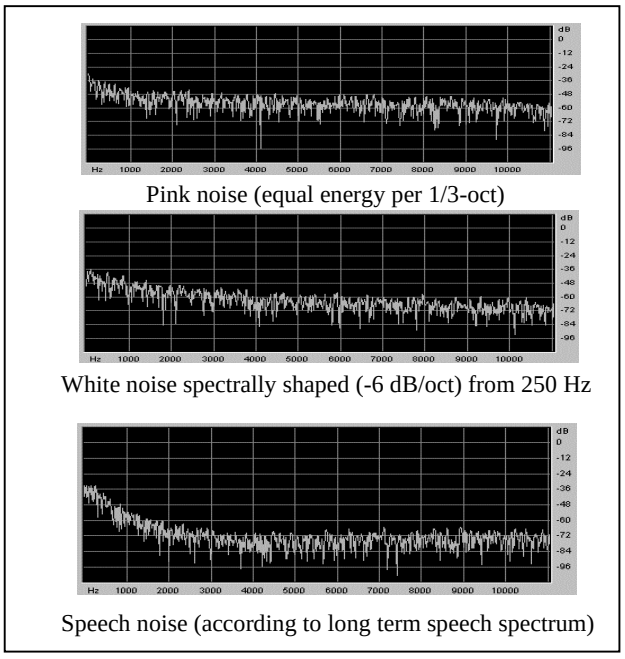

Figure 1: Noises used in the experiments. In the following sections the noise will be referred to *pinklvl*, *pinktilt* and *speech* from, respectively top to bottom in this figure. These noises have been extracted from the NATO RSG 10 database [12].

3. SPEAKER-DEPENDENT RECOGNITION RESULTS FOR THE LIVE EXPERIMENTS

Figure 2 summarizes the recognition accuracy obtained during the live experiments. As a dialogue manager presented the subjects with up to three candidates, the top 3 recognition results (meaning the right candidate was found among the top three candidates proposed by the system) have been plotted. It can be observed that recognition accuracy degrades a great deal when subjects are speaking while listening to noise at 85 dB SPL. There is a tendency to a decrease in recognition accuracy when the noise spectral tilt increases. However, in our experiments this difference is not significant. There is generally much more errors for female subjects than male subjects for all the conditions.

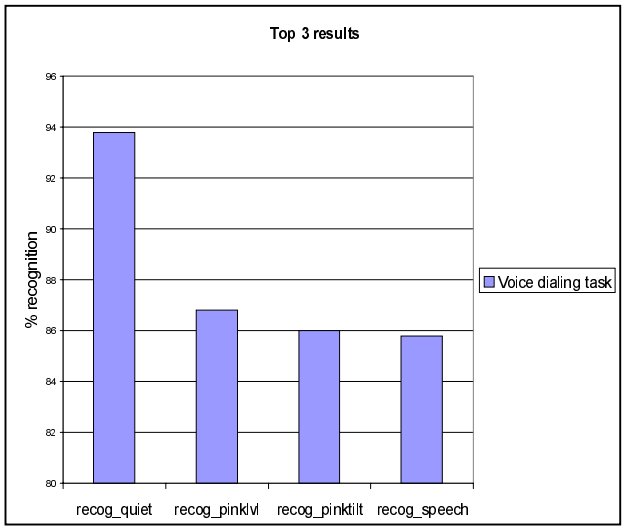

Figure 2: Live test recognition results in quiet environment and while subjects are listening to noise at 85 dB SPL for a voice dialing task.

4. ANALYSIS OF THE DATA RECORDED

We computed the signal to noise ratios (SNRs) on the data obtained during the recognition part of the experiments by averaging the variance of the energy in the speech and non-speech parts of the recorded signals (see Figure 3).

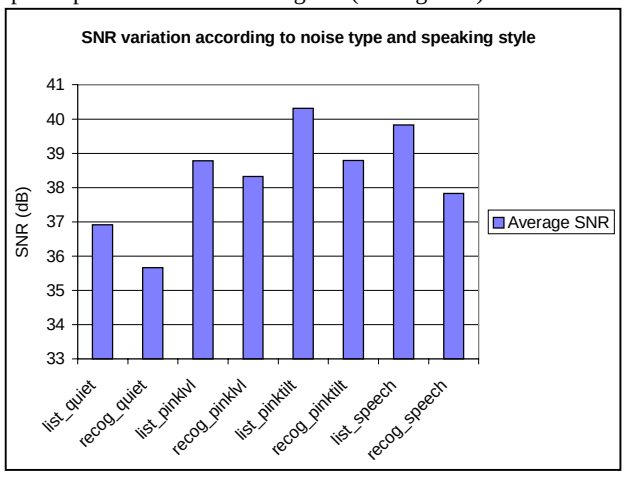

Figure 3: SNR averaged over the whole vocabulary and all the speakers for the different recorded conditions.

In communication with the voice dialing device there is a clear decrease of the SNR (up to 2 dB for the speech noise) as compared to when the subjects were reading a list. This is a very interesting result which supports the hypothesis that people produce speech differently when they do not perform a task. It also points out that studies of the Lombard reflex where data has been recorded while subjects are reading a list do not accurately represent the real conditions. Figure 4, which shows the average phrase duration over the whole vocabulary for male and female speakers, further emphasizes this point.

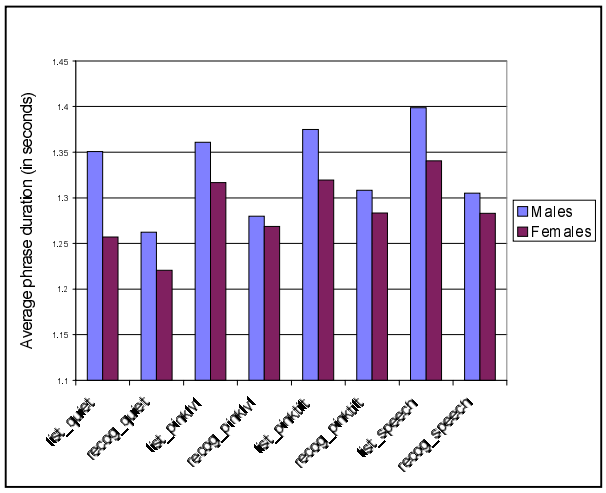

Figure 4: Average phrase duration over the whole vocabulary and all the recorded speakers for the different recorded conditions.

It is interesting to note that phrase duration decreases for all the conditions when subjects are speaking to the recognition device as compared to reading a list. This result is valid for both male and female speakers. All the subjects involved in the experiments exhibited this tendency. Moreover, duration tends to increase with the noise spectral tilt.

We also extracted pitch information (with Entropics ESPS tools) for all the vowels and diphtongs in the corpus ("ey", "ih", "iy", "ow", "oy", "uh", "uw"). When encountered, the voiceless parts of the segments were discarded. The pitch was computed as an average pitch across the whole vowel. Figure 5 shows the results for male and female speakers.

The pitch follows the same tendencies as the other parameters studied. As it has already been reported there is a pitch increase when the subjects are producing speech in noise. This result applies both when subjects are in a communication situation or are reading a list. When the noise spectral tilt increases, the pitch tends to decrease. The difference in pitch between the different conditions was noticeable to a linguist who listened to all the speech files.

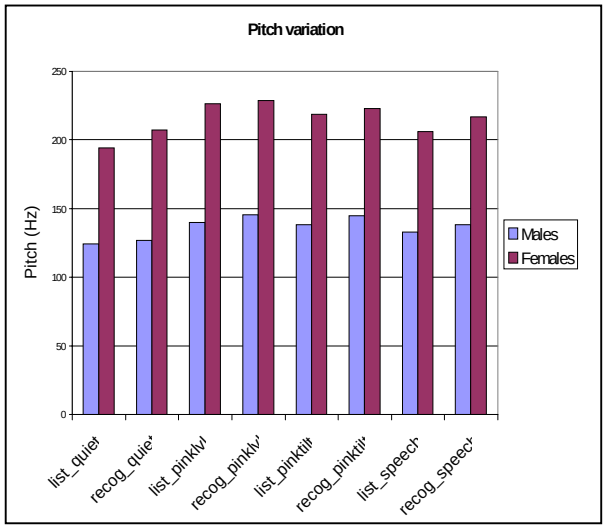

Figure 5: Averaged pitch values (in Hz) for all the vowels of the corpus and male and female speakers for all the recorded conditions.

5. OFF-LINE SPEAKER-INDEPENDENT RECOGNITION EXPERIMENTS

To evaluate how the acoustic changes reported in the previous section influence speaker-independent recognition we embedded the 50 words of the vocabulary in a larger vocabulary of 662 phrases (to emphasize the differences) and performed speaker-independent recognition experiments with an HMM-based flexible vocabulary recognizer using context-dependent models obtained after state-based decision tree clustering. One transcription per word was used. The experiments were run with three beam sizes: 0 for full search, 1.00 E-84 for a large beam, and 1.00 E-44 for a medium size beam. Results averaged over the 10 speakers are plotted in Figure 6.

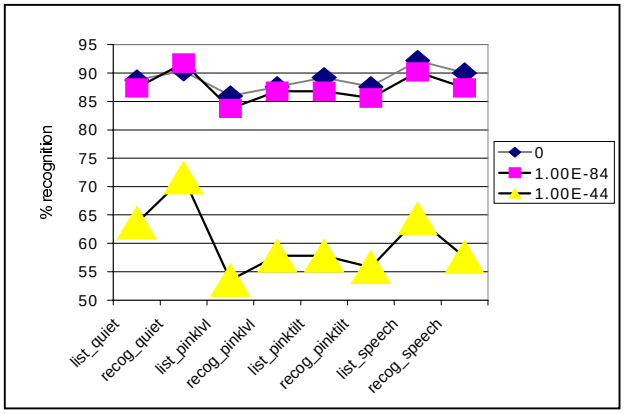

Figure 6: Speaker-independent recognition results across the conditions for the 50 phrases embedded in a 662 phrase vocabulary.

It is interesting to note that, in quiet, recognition accuracy is improved when subjects are talking to the device as compared to reading a list. One possible hypothesis is that the emphasis on

communication may help the user to speak more clearly. When subjects are reading a list, as noise spectral tilt increases recognition accuracy also increases. This phenomenon is less obvious when subjects are speaking to the voice recognition device.

At the recognition level, we observed that the type of noise affect speakers differently. This is especially true when subjects are in communication with the recognition device.

6. DISCUSSION

The above results highlight a very important point: *it is very important to study data representing realistic conditions*. The Lombard reflex is induced by noise. However, the modification of speech production in presence of noise is strongly influenced by the desire of the speakers to communicate. We presented results supporting very clearly that there is a definite influence of the communication factor on several acoustic correlates of the Lombard reflex and automatic speech recognition. The frequency distribution of the noise (in this study the noise spectral tilt) is also a factor to consider when studying the Lombard reflex. While the incidence of the noise on the Lombard reflex may be marginal for some acoustic correlates, we found that the influence of the noise spectral tilt on the user speech can be seen at the recognition level. As recognition accuracy is a global measure, the fact that acoustic changes occurring in the different noise conditions translate at the recognition level suggests that these differences are significant. New insights about the Lombard reflex have to be gained by studying databases when subjects are in communication with a device or performing a task. This will bring us closer from our final goal which is to improve speech recognition in noise. As speech recognition technology is progressing it is possible, as shown in this paper, to use the speech recognizers that we develop to record realistic data. When reading a list, users tend to over emphasize the changes in speech production, without directing these changes towards a specific goal. This creates an artificial situation.

7. CONCLUSIONS

In this paper we presented a procedure to record a Lombard database in realistic conditions using an automatic speech recognizer in the context of a voice dialing task. Different types of noise differing mainly by their spectral tilt were studied. The recorded data was analyzed and further off-line speech recognition experiments were performed.

When comparing data between subjects reading a list or in communication with a speech recognition device, SNRs and acoustic correlates of the Lombard reflex, such as duration and pitch, are affected by the communication factor. Speaker-independent recognition experiments performed off-line confirmed this hypothesis.

The noise type is another factor that influences the Lombard reflex. The variability observed between subjects seems to indicate that depending on the subjects and how they can cope with the type of noise, the Lombard reflex is been affected. The results obtained in this paper support this hypothesis.

Future studies will concentrate on a more detailed analysis of the data recorded and on the use of this data to benchmark new robust algorithms for dealing with the Lombard reflex.

8. ACKNOWLEDMENTS

The authors would like to acknowledge Moham Sondhi from Bell Laboratories, Lucent Technologies for suggesting the use of the open-ear headphones Sennheiser HD 580 to record Lombard speech.

9. REFERENCES

1. Junqua J-C. and Haton J-P. “Robustness in Automatic Speech Recognition”, Kluwer Academic Publishers, 1996

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. “Analysis and Compensation of Stressed and Noisy Speech with Application to Robust Automatic Recognition” Ph.D. thesis, 1988, Georgia Institute of Technology.

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. Fairbanks G. “Systematic Research in Experimental Phonetics. A Theory of the Speech Mechanism as a Servosystem”, J. Speech and Hearing Research, 1954, Vol. 19, pp.133-139.

9. Lombard E. “Le Signe de l’Elévation de la Voix”, Ann. Maladies Oreille, Larynx, Nez, Pharynx, 1911, Vol. 37, pp. 101-119.

10. Mokbel C. “ Reconnaissance de la Parole dans le Bruit: Bruitage/Débruitage”, Ph.D. Thesis, 1992, Ecole Nationale Supérieure des Télécommunications.

11. Noyes J. and Baber C. “Speech Recognition in Adverse Environments: The Role of Human Mediation” *ESCA/NATO Workshop on Speech Under Stress*, 1995, pp. 17–20.

12. Steeneken H. and Geurtsen F. “Description of the RSG-10 Noise Database”, Technical Report, 1990, TNO Institute for Perception