SPEECH PRODUCTION IN NOISY ENVIRONMENTS AND THE EFFECT ON AUTOMATIC SPEECH RECOGNITION

Panikos Heracleous[a,c], *Miki Sato*[b], *Carlos T. Ishi*[b], *Hiroshi Ishiguro*[a,c] *& Norihiro Hagita*[b]

[a]Hiroshi Ishiguro Laboratory, ATR, Japan;
[b]Intelligent Robotics and Communication Laboratories, ATR, Japan;
[c]Japan Science and Technology Agency, CREST, Japan
panikos@atr.jp

ABSTRACT

Speech is bimodal in nature and includes the audio and visual modalities. In addition to acoustic speech perception, speech can be also perceived using visual information provided by the mouth/face (i.e., automatic lipreading). In this study, the visual speech production in noisy environments is investigated. The authors show that the Lombard effect plays an important role not only in audio speech but also in visual speech production. Experimental results show that when visual speech is produced in noisy environments, the visual parameters of the mouth/face change. As a result, the performance of a visual speech recognizer decreases.

Keywords: speech, noisy environments, Lombard effect, lipreading

1. INTRODUCTION

In noisy environments, the talker increases the intelligibility of his/her speech [5], and, during this process, several characteristics of speech change (the Lombard effect) [1]. As a result, the performance of an automatic speech recognizer operating in a noisy environment decreases not only because of the noise contamination but also because of these modifications [3].

Although many studies have addressed the problem of the Lombard reflex in audio-only automatic speech recognition, only a few studies have addressed this issue with reference to automatic visual speech recognition. In [4], audiovisual speech recognition experiments using noisy and Lombard data were presented. In this study, it was also briefly mentioned that the Lombard effect is present not only in the audio channel but also in the visual channel, and a few results were also presented. In [2], the changes that occur in the visual correlates of speech articulation when speech is produced in noisy environments were considered. In this study, results were presented showing visual differences in the lip/mouth sector when speech was produced in a noisy environment or when Lombard speech was used. However, in this study, analysis and experimental results related to visual speech recognition were not reported.

In this study, the authors comprehensively analysed the visual Lombard effect phenomenon with respect to automatic visual speech recognition and showed significant progress compared to the previously limited studies. Specifically, continuous phoneme recognition experiments were conducted in Japanese, using data from several speakers. Further, a method based on adaptation was applied in order to address the problem of the Lombard effect in visual speech recognition.

2. ACOUSTIC LOMBARD REFLEX

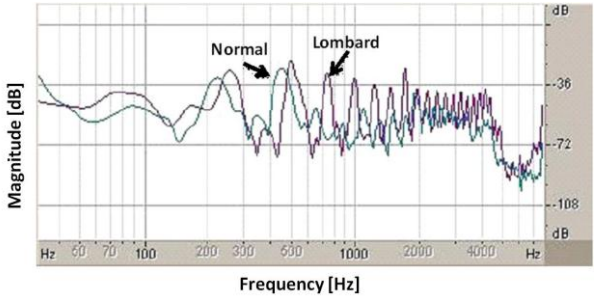
When speech is produced in noisy environments, the speech production process is modified, leading to the Lombard reflex. Specifically, due to the reduced auditory feedback, the talker attempts to increase the intelligibility of his/her speech. During this process, several characteristics of speech change. In particular, the intensity of speech increases, the fundamental frequency (F0) and formants shift, the durations of vowels in- crease, and the spectral tilt changes. Because of these modifications, the performance of a speech recognizer decreases.

One way to investigate the effect of the Lombard reflex is to analyze clean speech uttered while the speaker is listening to noise through headphones or earphones (i.e., Lombard speech). Even though Lombard speech does not contain any noise components, modifications in speech characteristics can be realized.

Figure 1 shows the power spectrum of a normal clean word and a Lombard word recorded while listening to office noise through headphones at 75

dB(A). The example clearly illustrates the modifications leading to the Lombard reflex: power is increased, formants are shifted, and spectral tilt is changed. These differences in the spectra cause feature distortions (e.g., distortions in the Mel-Frequency Cepstral Coefficients [MFCC]), and therefore, acoustic models trained using clean speech fail to correctly match the speech affected by the Lombard reflex.

Figure 1: Power spectrum of a normal clean and a Lombard utterance.



3. METHODOLOGY

3.1. Data and statistical modeling

For the experiments, three speakers (one male and two females) were instructed to read out sentences from the JNAS database. To obtain Lombard speech, the speakers listened to babble noises at 70, 75, and 80 dB(A) while uttering the sentences (i.e., Lombard data).

The data used were 400 continuous Japanese sentences [i.e., 250 clean utterances, 50 Lombard utterances with 70 dB(A), 50 Lombard utterances with 75 dB(A), and 50 Lombard utterances with 80 dB(A)]. Forty-three context-independent, hidden Markov models (HMMs) [6] trained using data from each speaker. Each HMM state was modeled with a mixture of 16 Gaussian components. The number of Gaussians was selected experimentally to obtain the highest accuracy. For training clean HMMs (i.e., trained with speech recorded in the clean environment), 15,706 phonemes were used, whereas 2,806 phonemes from each speaker of each noise level were used for testing. The acoustic parameter vectors were of length 36 (12 MFCC, 12 Δ MFCC, and 12 ΔΔ MFCC).

3.2. Visual parameter extraction

For lip-parameter extraction, the OKAO Vision commercial tool of the OMRON Corporation was used. Details concerning the methods applied for using the specific tool can be found in the study [8]. The OKAO Vision system carries out real-time detection and tracking of the face, mouth, and eyes, and each time frame provides the x-y coordinates of 38 points at a rate of 30 Hz. Using the 38 points provided, six lip parameters —along with their first- and second-order derivatives— were computed as follows: width (W), outer perimeter (C1), inner perimeter (C2), area (A), outer height (h1), and inner height (h2). Figure 2 shows the lip parameters used in this study.

Figure 2: Lip parameters used as features in the statistical modeling.

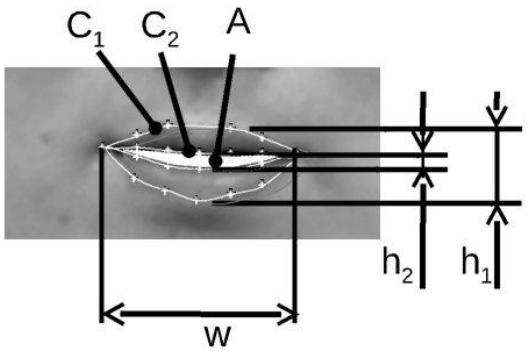
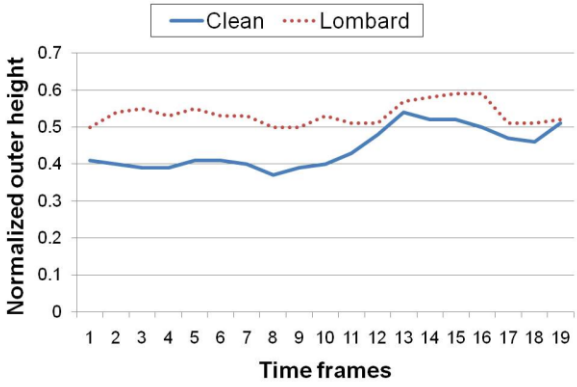


Figure 3: Normalized outer height in the case of a clean and a Lombard word.



To correct for the speaker–camera distance and the pose of the head, the lip features were normalized by dividing them with the Euclidean distance computed from the midpoint between the eyes and the upper lip, which does not move much during speech production. The visual signal was recorded at the rate of 30 Hz, in synchrony with the audio signal. A 25-ms window that shifted every 10 ms was used for extraction of the acoustic parameters. To obtain the same number of visual and audio samples, the visual samples were also interpolated before fusion was carried out.

4. EXPERIMENTS

4.1. Analysis of the visual parameters

Figure 3 shows the normalized outer height in the case of a Lombard utterance and a clean Japanese utterance. As is shown, in the case of the Lombard sentence, larger values are observed. Since these values are used as features in the statistical modeling, it is expected that differences in recognition rates will occur.

Figure 4 shows the density function of the outer height computed by the Kernel density estimation using the fast Fourier transform [7, 9] in the case of the Japanese male speaker. The figure clearly shows the differences when using clean speech, Lombard speech at the 70 dB(A) noise level, and Lombard speech at the 80 dB(A) noise level. By increasing the noise level, the probability of observing higher values in the outer height further increases.

Figure 4: Density functions in the clean and Lombard cases.

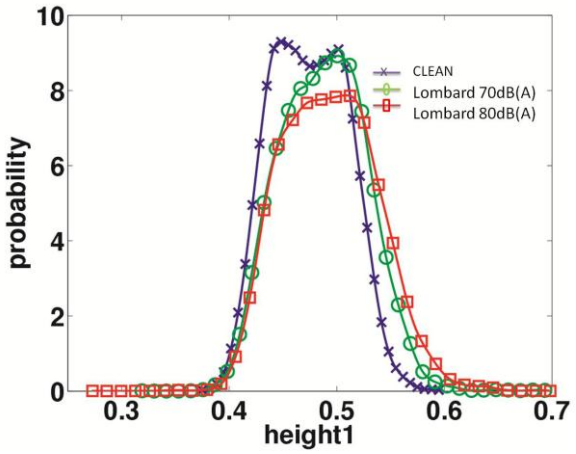


Table 1: Mean values of the visual parameters in the case of clean and Lombard speech.

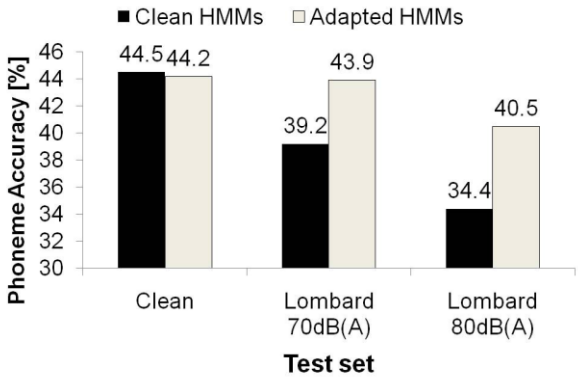
Parameter	*Visual Speech Data*		
	Clean	70dB(A)	80dB(A)
Outer height	0.474	0.488	0.492
Inner height	0.076	0.093	0.100
Width	0.753	0.773	0.777
Area	0.207	0.224	0.230
Outer perimeter	1.699	1.759	1.779
Inner perimeter	1.528	1.582	1.596

Table 1 shows the mean values of the normalized lip features over all test data in the case of Japanese clean speech and Lombard speech. In all cases, the parameters increase while using Lombard speech. The results also show that, as the noise level increases, larger differences can be observed.

4.2. Visual speech automatic recognition

Figure 5 shows the results obtained when Japanese multi-speaker automatic visual experiments were conducted. In this case, all the training data of the three speakers were used to train a common HMM set. The figure shows the effect of the Lombard reflex in visual speech recognition in Japanese continuous phoneme recognition. Using clean test data, the phoneme accuracy was 44.5%. When the test data comprised Lombard speech of 70 dB(A), the phoneme accuracy de- creased to 39.2%. The phoneme accuracy was further decreased upon increasing the noise level. Using the test Lombard speech of 80 dB(A), the phoneme accuracy was only 34.4%. The results show that, as the noise level increases, the phoneme accuracy further decreases.

Figure 5: Visual speech automatic recognition using clean HMMs and adapted HMMs.



To deal with the decreases of the phoneme accuracy due of the visual Lombard effect, the clean visual HMMs were adapted to the Lombard effect using Maximum Likelihood Linear Regression (MLLR) adaptation. MLLR is a method for speaker adaptation. In this study, however, MLLR was used to adapt the clean visual models to visual Lombard models. Specifically, 50 sentences from each speaker recorded at 75 dB(A) babble noise (i.e., different from the noise level of the Lombard test sets) were used as adaptation data, and MLLR was performed. Figure 5 also shows the results after MLLR was applied. As is shown, after model adaptation, the phoneme accuracies using Lombard data increased.

5. DISCUSSION

The current study focuses on the phenomenon of the Lombard effect with respect to automatic speech recognition. Although, speech cannot be perceived completely using visual information from mouth/lips alone, automatic visual speech recognition has applications in audiovisual speech recognition and in lip synthesis. It is important, therefore, to analyze the behaviour of automatic lip-reading also in adverse environments. In many audiovisual systems, audio speech is recorded in laboratory environments under relatively clean conditions. To better match the noisy testing conditions, artificially noisy training data are created by superimposing noise onto the clean training data and re-training the system. This is reasonable when artificial data are used for testing. In real applications, however, Lombard effect also appears. For robust audiovisual speech recognition in real environments the visual Lombard effect should also be considered.

6. CONCLUSIONS

In this study, the results obtained show that recognition rates decrease when visual speech recorded in a noisy environment (i.e. Lombard visual speech) is tested. In the case of a visual speech recognition system operating in a real noisy environment, further improvements in the recognition rates are achieved if the visual Lombard effect is also considered in the statistical model training.

7. ACKNOWLEDGEMENTS

This work has been partially supported by JST CREST 'Studies on Cellphone-type Teleoperated Androids Transmitting Human Presence'.

8. REFERENCES

[1] Bond, Z.S., Moore, T.J. 1990. A note on loud and lombard speech. *Proc. of International Conference on Speech and Language Processing*, 969-972.

[2] Davis, C., Kim, J., Grauwinkel, K., Mixdorff, H. 2006. Lombard speech: Auditory (a), visual (v) and av effects. *Proc. of Speech Prosody*.

[3] Hansen, J.H.L. 1996. Analysis and compensation of speech under stress and noise for environmental robustness in speech recognition. *Speech Communication, Special Issue on Speech Under Stress* 20(2), 151-170.

[4] Huang, F.J., Chen, T. 2001. Consideration of lombard effect for speechreading. *IEEE Fourth Workshop on Multimedia Signal Processing*, 613-618.

[5] Lombard, A.E.1911. Le signe de l'elevation de la voix. *Ann. Maladies Oreille, Larynx, Nez, Pharynx* 37, 101-119.

[6] Rabiner, L.R. 1989. A tutorial on hidden markov models and selected applications in speech recognition. *Proc. of the IEEE* 77(2), 257-286.

[7] Silverman, B.W. 1982. Kernel density estimation using the fast fourier transform. *J. Roy. Statist. Soc. Ser. C: Appl. Statist* 31(1), 93-99.

[8] Su, Y., Ai, H., Lao, S. 2008. Real-time face alignment with tracking in video. *Proc. of ICIP*, 1632-1635.

[9] Vlassis, N., Motomura, Y. 2001. Efficient source adaptivity in independent component analysis. *IEEE Trans. Neural Networks* 12(3), 559-566.