The Influence of Lombard Effect on Speech Driven Intelligent Environments

Damjan Vlaj[1]**, Zdravko Kačič**[1]

[1]University of Maribor, Faculty of Electrical Engineering and Computer Science,
Koroška cesta 46, 2000 Maribor, Slovenia
E-mail: damjan.vlaj@um.si, zdravko.kacic@um.si

Abstract

The main aim of the paper is to present the possible presence and the influence of Lombard effect on automatic speech recognition in speech driven intelligent environments. Nowadays, the presence of Lombard effect can be expected in contemporary speech recognition applications in numerous application domains, which are parts of the intelligent environments. In order to present the influence of Lombard effect on automatic speech recognition, we used the Slovenian Lombard Speech Database. During the tests when no noise was present in the testing environment, but only played via speaker's headphones, the influence of Lombard effect on speech recognition can easily be determined. When the highest level of noise was present, the speech recognition accuracy on application words decreased in average by 3.28 %.

1 Introduction

Intelligent environments are the environments with the integrated information and communication systems which enable interactive communication between a human being and a machine. In this paper we would like to present a speech driven intelligent environment. What we have in mind using this term, are especially intelligent houses, where an intelligent environment helps people in everyday tasks. Voice driven control of house lighting, household appliances, audio-visual equipment, and household robots (robotic vacuum cleaner, robotic mop, serving robot, etc.) improves the functionality of such intelligent environments. If we want to use a voice driven control in such intelligent environments, there are no major problems to be expected at the start-up. However, the major problems occur when we want to turn them off, because they produce noise. Noise can be generated both by electronic devices and by people. If we know the origin and form of the noise, it can be removed from the speaker's speech signal very effectively. In the acoustically noisy environments, the speaker speaks more loudly, which leads to the changes of the speech signal, in comparison to the speech signal, when noise is not present. This psychological effect of speech produced in the presence of noise is called Lombard effect [1].

The Lombard effect was first mentioned more than one hundred years ago. In 1911 Etienne Lombard discovered the psychological effect of speech produced in the presence of noise [1]. The Lombard effect is a phenomenon in which speakers increase their vocal levels in the presence of a loud background noise and make several vocal changes in order to improve intelligibility of the speech signal.

We especially want to point out the influence of Lombard effect on automatic speech recognition in noisy environment. In the intelligent environments the speech signal for automatic speech recognition is captured with microphone arrays [2]. It is also possible to use close-taking microphone, but in this case the speaker must carry a microphone with himself all the time. In the past, the project was performed, where the objective was to develop voice driven interfaces for consumer applications [3]. However, in this project the authors did not consider the impact of Lombard effect on automatic speech recognition. The Lombard effect occurs in acoustically noisy environments. As we have already mentioned, noise can be removed from the speaker's speech signal. If we look at audio-visual equipment (television, radio, etc.), we know the origin of the noise, which can then be removed. We can perform various analyses on the clean speech signal in order to determine the influence of Lombard effect.

In the next section our intelligent environment is presented. In the Section 3 the changes of speech signal characteristics in noisy environments are presented. With the experiments we want to confirm the influence of Lombard effect on speech recognition especially in noisy conditions. In the Section 4, the experimental design for speech recognition will be presented. The results of the experiments and discussion will be given in the Section 5 and the conclusion will be drawn in the Section 6.

2 Intelligent environment

Our intention is to use a voice driven control of the TV device. There are no major problems to be expected at turning the TV device on. However, the major problems can occur when we want to control or turn the TV set off, because it produces various types of audio signals. These audio signals can be various forms of music, movie special effects, voice of the predictor of the television shows, etc. From the speaker's point of view, these audio signals represent noise. In such noisy conditions the accuracy of automatic speech recognition can be reduced, comparing to the accuracy in a noise-free environment. This happens especially in the situations, when we do not have any information about the noise. However, noise captured in our situation (from TV device), can be removed from the speaker's speech signal, because we know its origin and form. For the noise reduction algorithm spectral subtraction based on minimum statistics was used, which was basically presented in [4] and its implementation was explained in [5]. Table 1 presents values of Signal to Noise Ratio (SNR) for various noisy environments before and after noise reduction.

However, in the acoustically noisy environments another problem occurs. In such environments the speakers speak louder, which causes them to increase their vocal levels and make several vocal changes in order to improve intelligibility of the speech signal. As we already mentioned, this phenomenon of speech production in the presence of noise is called Lombard effect. In the next section we will present changes of the speech signal characteristics in presence of Lombard effect.

Table 1. Signal to Noise Ratio (SNR) calculated for various noisy environments.

SNR condition	Calculated SNR [dB]	
	Before noise reduction	After noise reduction
High	30.53	32.23
Middle	16.45	30.96
Low	9.76	25.76

3 Changes of speech signal in noisy environments

In this section, we will present changes of speech signal in noisy environments, which leads us to the Lombard effect. The analysis was made on words pronounced in the Slovenian language. The following three speech characteristics in presence of Lombard effect are presented: value of pitch, phoneme duration, and frequency envelope with a shift of formant centre frequencies F1 and F2. The analysis of the characteristics was made on the pronounced word "prekliči" (in English this means "cancel") in three noisy environments.

3.1 Value of pitch

The analysis of the value of pitch was made on phoneme "e" and on the first phoneme "i" of the word "prekliči". Table 2 presents values of pitch measured for various noisy environments. With the presence of higher noise values (lower SNR) the speaker makes vocal changes in order to improve intelligibility of the speech signal. Looking at the obtained results the value of pitch increases, when the SNR condition decreases.

3.2 Phoneme duration

In this subsection the analysis of the duration of the phoneme "e" and the first phoneme "i" of the word "prekliči" was made. The duration analysis was made on vowels, because the difference is most obvious in the duration of vowels and not in the duration of consonants. The Table 3 presents the duration of phonemes measured for various noisy environments. For the analysed speech signal, the duration of phonemes in general increases, when the SNR condition decreases. The results indicate that the speaker, who pronounced analysed word, tended to increase the phoneme duration at higher level of background noise, but this does not seem to be as consistent as the increase of pitch.

Table 2. Values of pitch measured for various noisy environments.

SNR condition	Measured value of pitch [Hz]	
	For phoneme "e"	For phoneme "i"
High	106.66	111.88
Middle	126.98	138.41
Low	150.94	160.52

Table 3. Duration of phonemes measured for various noisy environments.

SNR condition	Duration of phonemes [ms]	
	For phoneme "e"	For phoneme "i"
High	45	70
Middle	101	123
Low	98	136

3.3 Frequency envelope

In this subsection the frequency envelopes in various noisy environments are presented. Figure 1 shows the frequency envelopes of the phoneme "e" and Figure 2 shows the frequency envelope of the first phoneme "i" of the word "prekliči". The increase of the first formant frequency is evident for both phonemes. Also an increase of energy in higher frequency range can be seen. Both features are known to occur in the speech, where Lombard effect is present. The changes of these features are less obvious for the words pronounced at the middle SNR condition than at the low SNR condition.

4 Experimental design

The experimental design was carried out on the Slovenian Lombard Speech Database [6, 7]. The Slovenian Lombard Speech Database is recorded in two recording sessions with at least one week pause between recordings. The recordings with and without presence of car or babble noise in the speaker's headphones were performed. Three recordings were made within one recording session: recordings of the signal without the presence of noise in the speaker's headphones, recordings of the signal with the presence of noise level of 80 dB SPL (Sound Pressure Level) in the speaker's headphones, and recordings of the signal with the presence of noise level of 95 dB SPL in the speaker's headphones.

For the training of acoustical models the speech material of the first session was used and for the testing the speech material of the second session was used. We also made cross experiments, so that we trained acoustical models on the second session and tested them

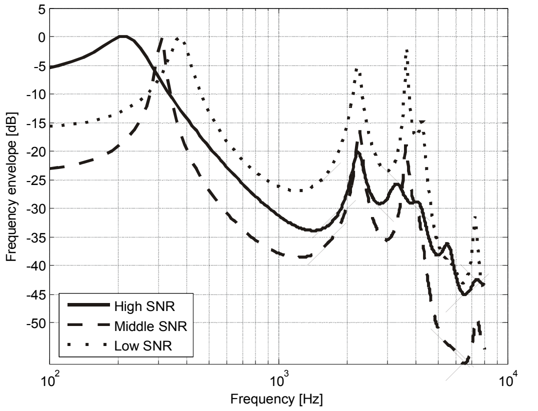

Figure 1. Frequency envelope of phoneme "e" of the word "prekliči" pronounced in various noisy environments.

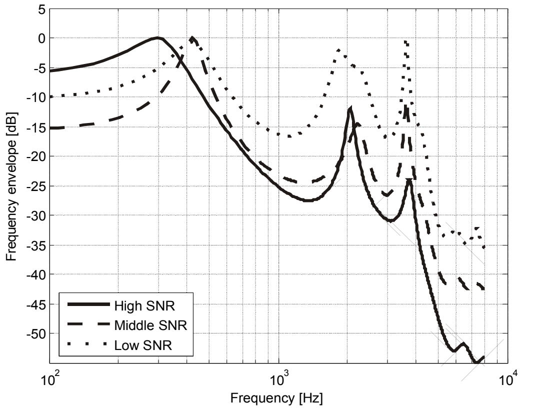

Figure 2. Frequency envelope of first phoneme "i" of the word "prekliči" pronounced in various noisy environments.

on the first session. For the training, the speech material recorded with hands free microphone was used. Here only the recordings of the signal without the presence of noise in the speaker's headphones were used. The experimental design for acoustic modelling was based on continuous Gaussian density Hidden Markov Models (HMM). The training was carried out using monophone acoustical models. The procedure for training of monophone acoustical models is presented in [6]. The Gaussian mixtures were increased by power of 2 up to 32 mixtures per state. Monophone acoustical models were trained on all eight corpuses of the Slovenian Lombard Speech Database [6, 7]. For this reason 2880 recorded files with 9474 pronounced words were used.

For the testing three scenarios were defined. For the first scenario, the recordings of noise mixed with speaker's speech that was played on speaker's headphones during recordings were used. For the second scenario, the same recordings were used, but here the noise reduction algorithm (spectral subtraction based on minimum statistics) was used [4, 5]. For the third scenario the speech material recorded with hands free microphone was used. Here the recordings of the signal with and without presence of noise in the speaker's headphones were used. In the tests, only babble noise was used, because this noise is most similar to the human speech in the tested intelligent environment. In the most cases TV devices emit human speech. The tests were made on application words corpus from the Slovenian Lombard Speech Database. The test set on application words contained 320 words. The testing of the monophone acoustical models was made after the second re-estimation of the HMMs with 32 Gaussian mixtures per state.

For the experimental design, we used Mel-cepstral coefficients and energy coefficient as features. We also used the first and the second derivative of the basic features. The features were created with the front-end using the basic distributed speech recognition standard from ETSI [8].

5 Results and discussion

In this section, the results obtained by the experiments will be presented. Figure 3 presents a chart, which shows the results of speech recognition accuracy tested on application words of the second session, and Figure 4 presents a chart, which shows the results of speech recognition accuracy tested on application words of the first session. We made cross experiments between two sessions, because we wanted to estimate the quality of acoustical models in both sessions. There are three groups with three speech recognition results presented on both charts. The first column in each group of the results presents the speech recognition accuracy when there was no noise played in the speaker's headphones. The second column presents the speech recognition accuracy when the babble noise was played in the speaker's headphones with the noise level of 80 dB SPL. This was a middle SNR condition in the speaker's environment. The third column presents the speech recognition accuracy when the babble noise in the speaker's headphones with the noise level of 95 dB SPL was present. This was a low SNR condition in the speaker's environment. At this point we would like to point out that in the recordings used for training of monophone acoustical models there is no noise present.

In the Figures 3 and 4, the three groups of testing environments can also be seen. The first testing environment presents noisy speech, where the speaker's speech was corrupted by the babble noise. The second testing environment presents noisy speech, where the noise reduction technique (spectral subtraction based on minimum statistics) was used [4, 5] in order to improve the characteristics of speech signal corrupted by the babble noise. The third testing environment presents clean speech, where no noise was present in the tested intelligent environment. In the first and second testing environment, we cannot determine the influence of Lombard effect on speech recognition, as the influence of environmental noise itself on speech recognition is too big. In the third testing environment, where no noise was present in the testing environment, but only in the speaker's headphones, the influence of Lombard effect on speech recognition can easily be determined.

As it can be seen from both figures, the speech recognition accuracy is very low if the speech recognition is carried out in noisy conditions when the SNR condition is low. When no noise reduction algorithm is used, the speech recognition in the intelligent environments is not applicable. When the

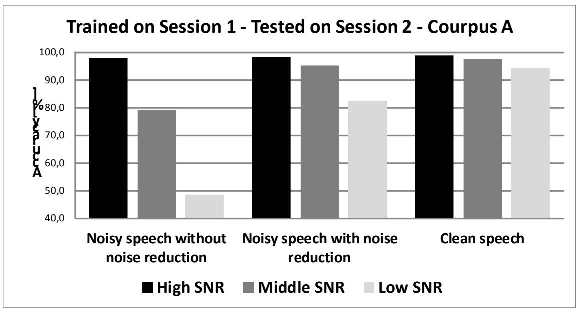

Figure 3. Speech recognition accuracy tested on application words (corpus A) of the second session.

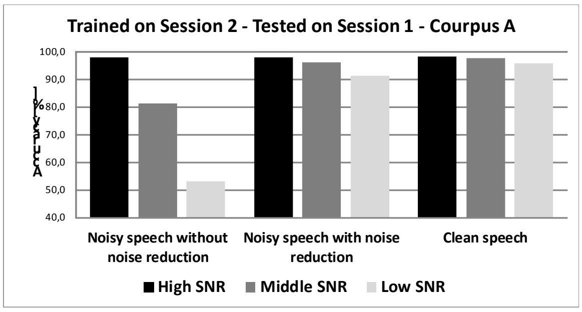

Figure 4. Speech recognition accuracy tested on application words (corpus A) of the first session.

noise reduction algorithm is used, the speech recognition accuracy can be improved a lot, (over the 90%) if the acoustical models are well trained and a noise reduction algorithm improves the quality of the speech signal.

Based on the speech recognition results for the third group of testing environments, we can conclude that the Lombard effect is present in the speaker's speech, which was pronounced with the noise present in the speaker’s headphones. When the noise level was 80 dB SPL, the speech recognition accuracy on application words decreased in average by 0.94 % for both testing sessions. When the noise level was 95 dB SPL, the speech recognition accuracy on application words decreased in average by additional 2.34 % for both testing sessions.

The speech recognition accuracy was almost always better when the monophone acoustical models were trained on first sessions and tested on second session. The reason for this could lay in a better trained monophone acoustical models on the first session or better acoustical environment in the second session of the Slovenian Lombard Speech Database. Should the second reason be the right one, it could be concluded that the speakers have adapted to the recording environment. Namely, when speakers recorded the second session, they had already known what to expect.

6 Conclusion

In this paper a speech driven intelligent environment and potential Lombard phenomena were presented and evaluated. Our intelligent environment is conceived as an intelligent room with the audio-visual equipment, which can be managed and controlled by automatic speech recognition. The main aim of the paper was to present the influence of Lombard effect on automatic speech recognition in speech driven intelligent environments. We made experiments to present the influence of Lombard effect on automatic speech recognition. In order to do so, we used the Slovenian Lombard Speech Database. The experiments were made with and without noise present in the testing environment. From the results it can be seen that speech recognition accuracy is very low if the speech recognition is carried out in noisy conditions with low SNR condition. In a noisy testing environment, we cannot determine the influence of Lombard effect on speech recognition, as the influence of environmental noise on speech recognition is too big. We can conclude that the Lombard effect is present in the speaker's speech, which was pronounced within the noisy environment. When the noise level in the speaker’s headphones was 80 dB SPL, the speech recognition accuracy on application words decreased in average by 0.94 %. When the noise level was increased from 80 dB SPL to 95 dB SPL, the speech recognition accuracy on application words decreased in average for additional 2.34 %.

References

[1] E. Lombard, Le signe de l’elevation de la voix, Annals maladiers oreille, Larynx, Nez, Pharynx, Vol. 37, pp. 101-119, 1911.

[2] J. Benesty, J. Chen, Y. Huang, Microphone Array Signal Processing, Springer Topics in Signal Processing, Vol. 1, Berlin Heidelberg, Springer Verlag, 250p, 2008.

[3] D. Iskra, B. Grosskopf, K. Marasek, H. Heuvel, F. Diehl, A. Kiessling, SPEECON - Speech Databases for Consumer Devices: Database Specification and Validation, Proceedings of Third International Conference on Language Resources and Evaluation – LREC'02, Las Palmas, Spain, pp. 329-333, 2002.

[4] R. Martin, Spectral subtraction based on minimum statistics. EUSIPCO'94 Proceedings. Edinburgh, Scotland, UK. pp. 1182-1185, 1994.

[5] B. Kotnik, D. Vlaj, B. Horvat, Efficient Noise Robust Feature Extraction Algorithms for Distributed Speech Recognition (DSR) Systems, International Journal of Speech Technology, Vol. 6, No. 3, pp. 205-219, 2003.

[6] D. Vlaj, Z. Kačič, The Influence of Lombard Effect on Speech Recognition, Speech Technologies, Rijeka: InTech. 432p, 2011.

[7] D. Vlaj, A. Zögling Markuš, M. Kos, Z. Kačič, Acquisition and Annotation of Slovenian Lombard Speech Database, Proceedings of the seventh Conference on International Language Resources and Evaluation – LREC'10, Valletta, Malta, pp. 595-600, 2010.

[8] ETSI standard document, ETSI ES 201 108 v1.1.1 - Speech Processing, Transmission and Quality aspects (STQ), Distributed speech recognition, Front-end feature extraction algorithm, Compression algorithm, Valbonne, France, 2000.