Acquisition and Annotation of Slovenian Lombard Speech Database

Damjan Vlaj, Aleksandra Zögling Markuš, Marko Kos, Zdravko Kačič

Faculty of Electrical Engineering and Computer Science, University of Maribor
Smetanova ulica 17, 2000 Maribor, Slovenia
damjan.vlaj@uni-mb.si, sandra.zogling@uni-mb.si, marko.kos@uni-mb.si, kacic@uni-mb.si

Abstract

This paper presents the acquisition and annotation of Slovenian Lombard Speech Database, the recording of which started in the year 2008. The database[1] was recorded at the University of Maribor, Slovenia. The goal of this paper is to describe the hardware platform used for the acquisition of speech material, recording scenarios and tools used for the annotation of Slovenian Lombard Speech Database. The database consists of recordings of 10 Slovenian native speakers. Five males and five females were recorded. Each speaker pronounced a set of eight corpuses in two recording sessions with at least one week pause between recordings. The structure of the corpus is similar to SpeechDat II database. Approximately 30 minutes of speech material per speaker and per session was recorded. The manual annotation of speech material is performed with the LombardSpeechLabel tool developed at the University of Maribor. The speech and annotation material was saved on 10 DVDs (one speaker on one DVD).

1. Introduction

The goal of this paper is to describe the hardware platform used for acquisition of speech material, to present recording scenarios and to present a tool used for the annotation of Slovenian Lombard Speech Database.

The origin of Lombard effect dates back to 1911 when Etienne Lombard (Lombard, 1911) discovered the psychological effect of speech produced in the presence of noise. 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.

The conducted research reported in the literature showed that Lombard speech is different from normal speech in a number of ways. The main changes of characteristics of Lombard speech can be seen in increase at voice level, fundamental frequency and vowel duration, and a shift in formant center frequencies for F1 and F2.

As written in (Bořil et al., 2006) speech databases recorded in real environments provide valuable material for speech recognition systems, but in case of louder backgrounds (car noise, babble noise …), it is difficult to analyze the signal, so that the analysis could confirm the presence of the Lombard effect. To confirm the presence of the Lombard effect in the signal, the speech signal that contains as less background noise as possible is required. Therefore Bořil et al. recorded a speech database in which they wanted to highlight the impact of the Lombard effect (Bořil et al., 2006). To further enhance such research efforts, the objective of recording the Slovenian Lombard Speech Database was to show the impact of the Lombard effect in the two different noise levels heard by the speaker, using two different noise types and recording speech in two recording sessions. With such database design we wanted to enable analysis that would show whether the increase of the noise level also increases the influence of the Lombard effect. We further wanted to enable analysis of Lombard effect consistency during two recording sessions, because at least one week pause between recording sessions was considered.

The paper is organized as follows: acquisition of raw audio material recorded in studio conditions is described in Section 2. Annotation of speech material and conversion of the audio material to the final format are presented in Section 3. The structure of Slovenian Lombard Speech Database is presented in Section 4. The results of the Lombard speech analysis are given in Section 5 and the conclusion is given in Section 6.

2. Acquisition of raw audio material

The Slovenian Lombard Speech Database was recorded in studio environment. Each speaker pronounced a set of eight corpuses in two recording sessions with at least one week pause between recordings. Approximately 30 minutes of speech material per speaker and per session was recorded.

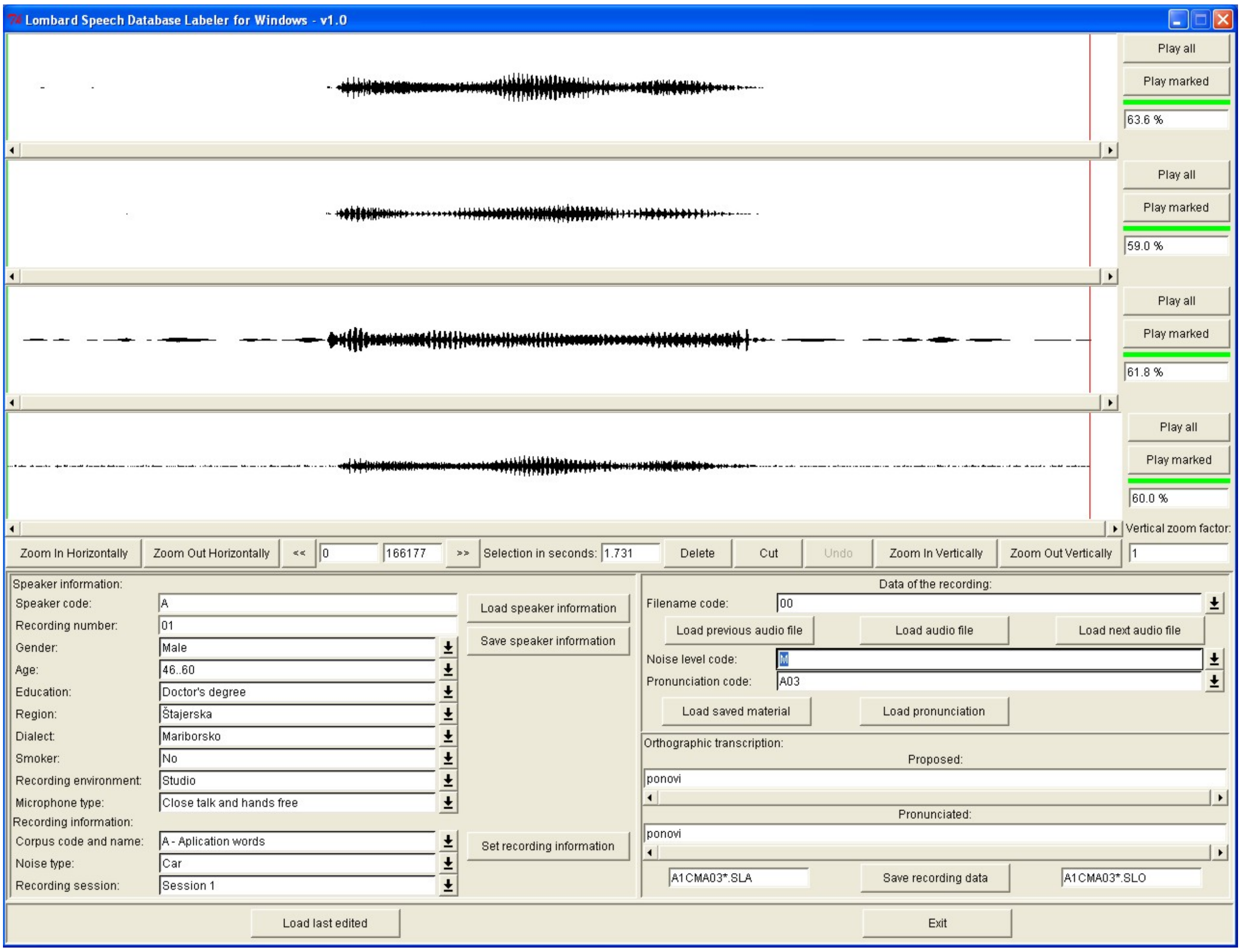
The recordings were performed using a hands-free microphone AKG C 3000 B, close talking microphone Shure Beta 53 and two channel electroglottograph EG2. Four channel recordings were performed:

- hands free microphone,
- close talking microphone,
- laryngograph and
- recordings of noise mixed with speaker's speech that was played on speaker's headphones during recordings.

The recording platform consisted of Audigy 4 PRO external audio card for 4 channel audio recording, Phonic MU244X mixer, and using 96 kHz sampling frequency, 24-bit linear quantization.

Two types of noises were used in recordings: babble and car noise. The noises were taken from the Aurora 2

[1] The owner of the database is SVOX.



Lombard Speech Database Labeler for Windows - v1.0

63.6 %

59.0 %

61.8 %

60.0 %

Selection in seconds: 1.731

Vertical zoom factor: 1

Speaker information:

Speaker code: A

Recording number: 01

Gender: Male

Age: 46..60

Education: Doctor's degree

Region: Štajerska

Dialect: Mariborsko

Smoker: No

Recording environment: Studio

Microphone type: Close talk and hands free

Recording information:

Corpus code and name: A - Aplication words

Noise type: Car

Recording session: Session 1

Data of the recording:

Filename code: 00

Noise level code: M

Pronunciation code: A03

Orthographic transcription:

Proposed: ponovi

Pronunciated: ponovi

A1CMA03*.SLA

A1CMA03*.SLO

Figure 1: LombardSpeechLabel tool for manual annotation of speech material.

database (Hirsch & Pearce, 2000) and were normalized. The noises were played to speaker's headphones AKG K271.

At the beginning of each recording the level of the reproduced background noise was adjusted according to the scheme proposed in (Bořil et al., 2006). The required noise level was adjusted by setting the corresponding effective voltage of the sound card open circuit VRMS OL. Noise levels of 80 dB SPL[2] and 95 dB SPL at a virtual distance of 1–3 meters were used for the Lombard speech recordings.

Three recordings of all corpuses were made within one recording session:

- without noise (reference recording),
- at 80 dB SPL and
- at 95 dB SPL.

A short pause was made between recordings of items of particular corpus (word, number, number string, sentence) to allow speaker's recovery. After the complete corpus was recorded a longer pause was made to allow for speaker's recovery.

There was an interaction between the "Lombard" speaker and a listener. The listener heard the attenuated speech mixed with non attenuated noise, evaluated the intelligibility and reacted accordingly. The reaction of the listener was mediated to the speaker by means of message displayed on the LCD display, where the speaker was notified that the pronunciation was intelligible or she/he was asked to repeat the pronunciation as it was not intelligible enough.

3. Annotation of speech material

The manual annotation of speech material is performed by the LombardSpeechLabel tool (see Figure 1) developed at the University of Maribor. The program tool is written in the Tcl/Tk/Tix language, which is suitable for visual programming. It was developed on the Microsoft Windows platform and can be incorporated into other operating system platforms with small modifications.

The LombardSpeechLabel tool window is divided into three fields. The upper field contains four waveform views (hands free microphone, close talking microphone, laryngograph and recordings played on speaker's headphones) of the signal that have been captured during recording of the database. By clicking the buttons on the right hand side of the upper field, each signal can be played individually. The bottom of the tool window is divided into two parts. On the left hand side the information about the speaker and the recording is given. On the right hand side, the additional data of the recording

[2] SPL is abbreviation for Sound Pressure Level.

and the orthographic transcription are presented.
The conversion of the audio material to the final format, which was set to 96 kHz sampling frequency, 16-bit linear quantization is also made with the LombardSpeechLabel tool.

4. The structure of the database

The Slovenian Lombard Speech Database consists of recordings of 10 Slovenian native speakers. Five males and five females were recorded. As we already mentioned, each speaker pronounced a set of eight corpuses in two recording sessions with at least one week pause between recordings. The corpus's structure is similar to SpeechDat II database (Kaiser & Kačič, 1997). In the following subsections more information about the database will be given.

4.1 Audio and label file format

Audio files are stored as sequences of 16-bit linear quantization at the sampling frequency of 96 kHz. They are saved in Intel format. Each prompted utterance is stored in a separate file. Each speech file has an accompanying SAM label file with UTF-8 symbols.

4.2 File nomenclature

File names follow the ISO 9660 file name conventions (8 plus 3 characters) according to the main CD ROM standard. Owing to the large amounts of audio material, the data were stored on a DVD-ROM media.

A	Speaker code (A-Z)
S	Session code (1-9) – used only 1 and 2
T	Code of the noise type: • R: without noise • C: Car noise • B: Babble noise
R	Code of the recording: • N: recording of the reference signal without presence of noise • L: recording of the signal without presence of noise • M: recording of the signal with presence of noise level of 80 dB SPL • H: recording of the signal with presence of noise level of 95 dB SPL
NNN	Code of the corpus (A00 – Z99): A – application words, B – connected digits, D – dates, I – isolated digits, N – natural numbers, S – phonetically rich sentences, T – times, W – phonetically rich words
C	Code of the recording channel: • 1: hands-free microphone • 2: close talk microphone • 3: signal captured by laryngograph • 4: signal in headphones that was heard by a speaker
LL	Two letter ISO 639 language code
F	File type code O=Orthographic label file, A=audio speech file

Table 1: Description of file nomenclature.

<database>	Defined as: <name><language code> i.e. LOMBSPSL Where: <name> is LOMBSP indicating Lombard Speech <LL> is the ISO 2-letters code SL for Slovenian
<speaker>	Defined as: SPK_<a> Where <a> is a progressive letter from A to Z. This letter is the same as the first letter used in file names (see section 4.2).
<session>	Defined as: SES_<s> Where <s> is a progressive number in the range 1 to 9. This number is the same as the second number used in file names (see section 4.2).
<condition>	Tree types of conditions are defined: • REF: recording of the reference signal without presence of noise, • CAR: recording of the signal with presence of car noise and • BABBLE: recording of the signal with presence of babble noise
<corpus>	Defined as: CORPUS_<c> Where <c> is a letter for one of corpus defined: A – application words, B – connected digits, D – dates, I – isolated digits, N – natural numbers, S – phonetically rich sentences, T – times, W – phonetically rich words

Table 2: Lombard speech database directory structure.

The following template for file nomenclature is used:

```
A S T R NNN C. LL F
```

The file nomenclature is described in Table 1.

4.3 Directory structure

The directory structure is set so that each speaker is located on his own DVD-ROM volume. Each speaker has two sessions. In each session the reference condition and two noise conditions are included. Each condition includes eight corpses. The following five levels directory structure is defined:

```
\<database>
    \<speaker>
        \<session>
            \<condition>
                \<corpus>
```

The Lombard speech database directory structure is presented in Table 2.

4.4 Corpus code definition

As it is useful for users to clearly identify the speech file contents by looking at the filename, we have specified the corpus code to support one letter corpus identifier and two numbers identifier. The corpus code definition is described in Table 3.

Corpus identifier	Item identifier	Corpus contents
A	00-29	application words (30 words)
B	00-04	connected digits (10 digits sequence pronounced 5 times)
D	00-04	dates (5 dates)
I	00-11	isolated digits (12 digits)
N	00-04	natural numbers (5 numbers)
S	00-29	phonetically rich sentences (30 sentences)
T	00-06	times (7 times)
W	00-49	phonetically rich words (50 words)

Table 3: Corpus code definition.

5. Results of the Lombard speech analysis

Two main interpretations of the Lombard effect have been proposed. The first argues that the effect is a physiological audio-phonatory reflex (Lombard, 1911); the second that Lombard changes are motivated by compensation on the part of the speaker for decreased intelligibility (Lane & Tranel, 1971). Some authors have also argued that both mechanisms may contribute to the changes made by the speaker in noisy environments (Junqua, 1993).

Detailed surveys of the literature on the Lombard effect phenomenon were made in (Lane & Tranel, 1971) and more recently in (Junqua, 1996). The conducted research showed that Lombard speech is different from normal speech in a number of ways. The main changes of characteristics of Lombard speech can be seen in increase in voice level, fundamental frequency and vowel duration, and a shift in formant center frequencies for F1 and F2. It was also reported in (Hanley & Steer, 1949) that speaking rate may be reduced when speech is produced in a noisy environment.

In this paper, we will make speech analysis for three Lombard speech characteristics: mean value of pitch, phoneme duration and frequency envelope.

In the detailed analysis the all three Lombard speech characteristics were measured for different voiced phonemes for the utterances of three words: "ustavi" (stop), "ponovi" (repeat) and "predhodni" (previous). In this paper only the selected results of Lombard speech analysis will be presented.

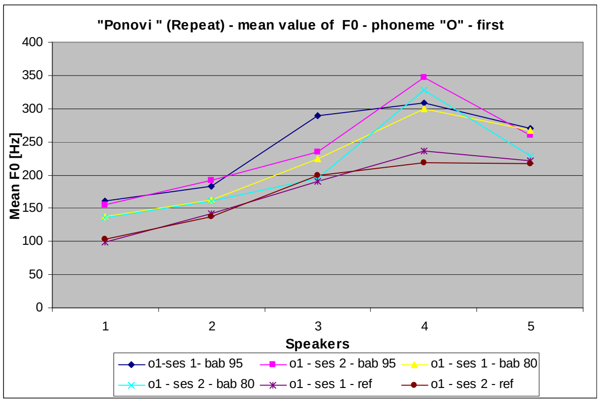


Figure 2: Mean pitch values of the first phoneme “O” of the word “ponovi” (Repeat) recorded at different noise levels and at babble background noise.

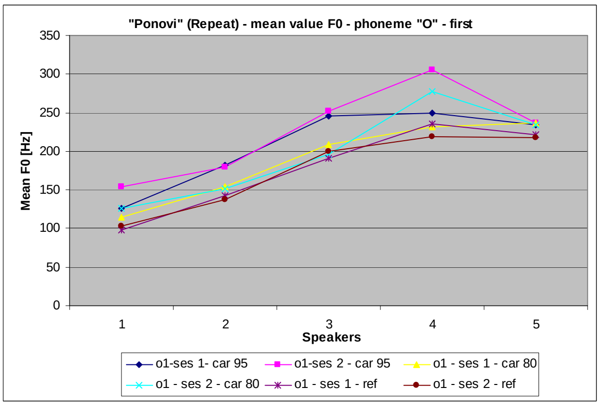


Figure 3: Mean pitch values of the first phoneme “O” of the word “ponovi” (Repeat) recorded at different noise levels and at car background noise.

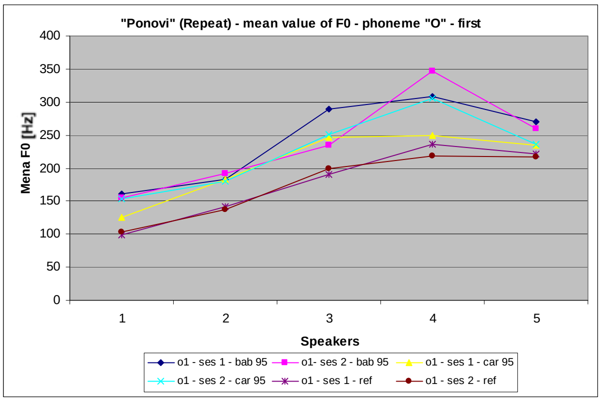


Figure 4: Mean pitch values of the first phoneme “O” of the word "ponovi" (Repeat) recorded at 95 dB SPL background noise and various noises.

5.1 Mean value of pitch

According to the literature the value of pitch increases in Lombard speech compared to normal speech. In this section the results of mean pitch values of the first phoneme "O" of the word "ponovi" (Repeat) will be presented. Figures 2 to 4 show the mean pitch values of voiced speech (vowel "O") for five speakers, for two sessions and two noise types. Speakers 1 and 2 were male speakers, whereas speakers 3 to 5 were female speakers.

Significant increase of pitch in first vowel "O" of the word "ponovi" (repeat) compared to reference pronunciations can be seen on Figures 2 and 3 for Lombard speech recorded under 95dB noise level for all five speakers. The increase can be observed in both recording sessions and for both noise types, although the extent varies among speakers. The increase is almost the same for the first, second and the fifth speaker and varies most for the third speaker in case of babble background noise. In case of car background noise the difference is bigger for the first and the forth speaker. For utterances recorded under 80 dB noise level the increase of pitch is significant in case of

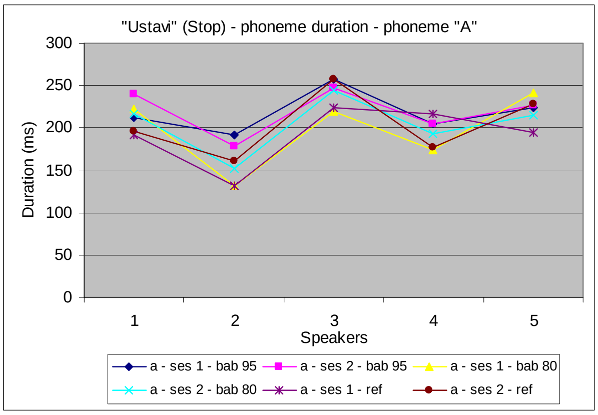


Figure 5: Duration of the phoneme "A" of the word "ustavi" (Stop) recorded at babble background noise and at different noise levels.

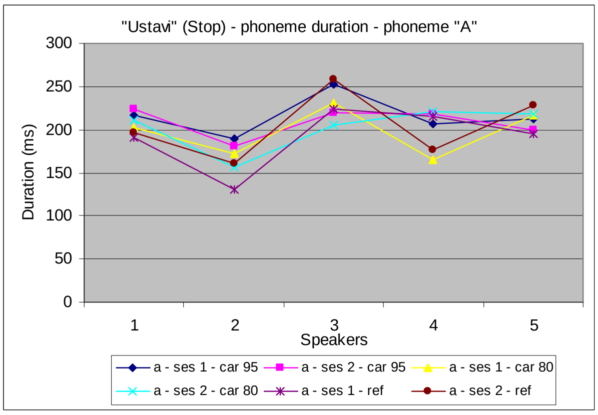


Figure 6: Duration of the phoneme "A" of the word "ustavi" (Stop) recorded at car background noise and at different noise levels.

babble noise (except for third speaker) but is less clear in case of car noise for most speakers.

Figure 4 shows that the greatest variability of pitch values for both types of noises in both recording sessions can be observed for speaker 4, whereas the values are the most consistent for speaker 2.

5.2 Phoneme duration

In this section the results of the duration of the vowel "A" of the word "ustavi" (stop) for all five speakers are presented. Figures 5 and 6 show the results of the analysis. It can be seen that the duration varies among speakers, but is more consistent per speaker regarding different recording sessions, background noise type and noise level. However, there is no clear distinction in phoneme duration concerning different recording sessions, background noise level or noise type. Figures 5 and 6 indicate that speakers tend to increase the phoneme duration at higher level of background noise, but this seems to be not as consistent as the increase of pitch.

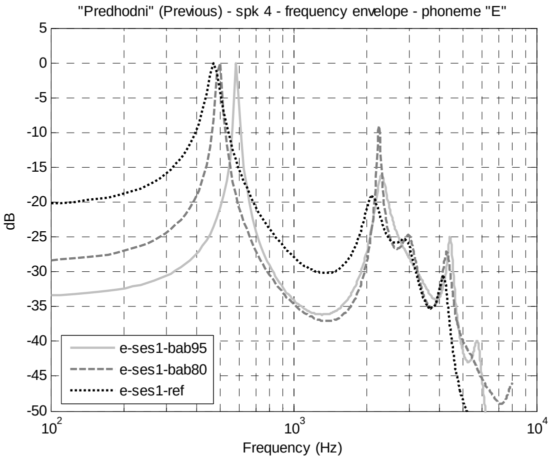


Figure 7: Frequency envelope of phoneme "E" of the word "Predhodni" (Previous) recorded at babble background noise and at different noise levels for female speaker (speaker 4) and for the first recording session.

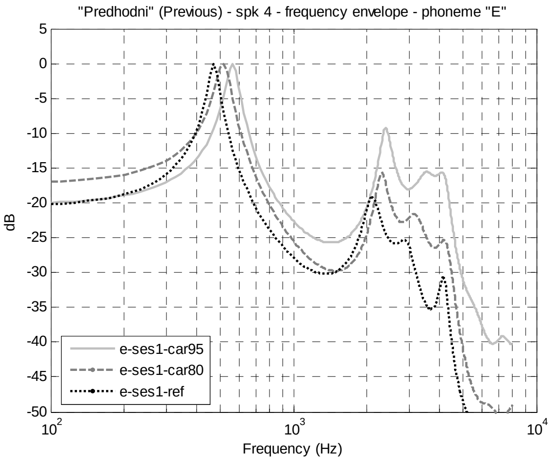


Figure 8: Frequency envelope of phoneme "E" of the word "Predhodni" (Previous) recorded at car background noise and at different noise levels for female speaker (speaker 4) and for the first recording session.

5.3 Frequency envelope

In this section the results of frequency envelope of phoneme "E" of the word "Predhodni" (Previous) recorded at different background noises and at different noise levels for female speaker (speaker 4) are presented. Figures 7 and 8 show these results of the analysis. The increase of the first formant frequency is evident for both background noise types. Also an increase of energy in higher frequency range can be seen. Both features are known to occur in Lombard speech. The changes of these features are less obvious for utterance uttered at 80 dB background noise.

6. Conclusion

In this paper the hardware platform used for acquisition of speech material, recording scenarios and tool used for the annotation of Slovenian Lombard Speech Database are presented. The database consists of recordings of 10 Slovenian native speakers. Five males and five females were recorded. Each speaker pronounced the set of eight corpuses in two recording sessions with at least one week pause between recordings. Approximately 30 minutes of speech material per speaker and per session was recorded.. The speech analysis was performed for three Lombard speech characteristics: mean value of pitch, phoneme duration and frequency envelope. The results of the analysis preformed indicate that the recorded speech encompasses several speech features generally known to be present in Lombard speech. Form this it can be concluded that the recorded speech database represents an appropriate speech material for further research work.

7. References

Lombard E. (1911). Le signe de l'elevation de la voix, *Annals maladiers oreille, Larynx, Nez, Pharynx*, vol. 37, pp. 101 – 119.

Junqua J-C. (1996). The influence of acoustics on speech production: A noise-induced stress phenornenon known as the Lombard reflex, *Speech Communication*, 20, pp. 13 – 22.

Bořil H., Bořil T. & Pollák P. (2006). "Methodology of Lombard speech database acquisition: Experiences with CLSD", In *Proceedings of the fifth Conference on Language Resources and Evaluation – LREC'06*, pp. 1644 – 1647.

Hirsch H. G. & Pearce D. (2000). The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions, In *Proceedings of the ISCA ITRW ASR'00*, Paris, France.

Kaiser J. & Kačič Z. (1997). SpeechDat Slovenian Database for the Fixed Telephone Network, University of Maribor, Maribor, Slovenia.

Lane H. & Tranel, B. (1971). The Lombard sign and the role of hearing in speech, *Journal of Speech and Hearing Research*, 14(4), pp. 677 – 709.

Junqua J-C. (1993). The Lombard reflex and its role on human listeners and automatic speech recognizers, *Journal of the Acoustical Society of America*, W(l), pp. 510 – 524.

Hanley T. & Steer M. (1949). Effect of level of distracting noise upon speaking rate, duration and intensity, *Journal of Speech and Hearing Disorders*, 14(4), pp, 363 – 368.