Lombard speech database for German language

Michał Sołoducha[1], Alexander Raake[2], Frank Kettler[3], Peter Voigt[4]

[1] *Technische Universität Ilmenau, 98693 Ilmenau, Germany, Email: michal.soloducha@tu-ilmenau.de*

[2] *Technische Universität Ilmenau, 98693 Ilmenau, Germany, Email: alexander.raake@tu-ilmenau.de*

[3] *HEAD acoustics GmbH, 52134 Herzogenrath, Germany, Email: frank.kettler@head-acoustics.de*

[4] *AVM GmbH, 10559 Berlin, Germany, Email: p.voigt@avm.de*

Abstract

In the scope of the presented work a German speech database has been created that will be published online for further reuse. Lombard speech was stimulated during recordings by presenting the overall 8 German native speakers (5 male, 3 female) with three different noise conditions via headphones, causing different degrees of the typical raised voice volume and specifically stressed speech (pitch, timing, timbre). The noise stimulus used during recordings is an artificially created babble speech consisting of multiple overlaid speech recordings, and is characterized by a good compromise between stationarity and naturalness. Due to the applied headphone presentation, the speech recordings do not include the noise stimuli, and may be mixed with different types of noises in the post-processing phase. Such a material can be used for a variety of subjective and instrumental tests addressing speech in noisy conditions.

Introduction

For speech communication, addressing our noisy environment has become more and more important in the last decades. This is mainly due to an increased usage of communication devices in mobility scenarios that are associated with interfering noises. This may imply problems in communication between people especially in the case of mediated conversations. To alleviate the problem of noise, a variety of signal processing algorithms have been developed in the past for noise cancellation (sending-side noise) or speech enhancement targeting, for example, an improved speech intelligibility (for sending and/or receiving side noise) [1, 2]. Moreover, there is a range of research work in the field of noise perception. That includes studies addressing e.g. intelligibility or quality of different telecommunication services.

Whenever speech in noise is addressed, the Lombard effect has to be taken into consideration [3]. Typically this effect results in a raised voice volume and stressed speech in terms of pitch, timbre and timing. This phenomenon helps humans to maintain good speech intelligibility in noisy environments while communicating to each other. This is the reason for including the effect in different kinds of studies addressing speech in noise [4]. There are standardized procedures to adjust the speech stimuli volume to a level which would be reached by a human in noisy environments [5]. This does not include, however, other modifications of a voice, i.e. spectral and temporal modifications typical during naturally produced Lombard speech. There is currently no model of Lombard speech including these factors so real Lombard speech recordings have to be used. Since no database of this type enabling narrowband, wideband and super-wideband speech-processing research was available to us, a corresponding Lombard speech database has been recorded in this work, which can be reused in different research contexts.

To produce a universal Lombard speech database, several compromises have to be taken into consideration. First of all, environmental noises have different levels. For example, in the standardized and publicly available background noise databases the levels range between 56 dBA for a living room and 81 dBA inside of an aircraft [6, 7]. Moreover, the levels above 85 dBA may be considered as harmful to human hearing with the exposure time 8 h (94 dBA for 1 h) [8]. Hence, it has been decided by the authors to consider only three noise conditions: silent environment, 55 dBA and 70 dBA which are considered to be representative of typical, not overly excessive levels which could be encountered in everyday situations. Noises at level of 55 dBA are considered by light but audible and a noise level of 70 dBA are already considerably loud noises which may have significant influence on human communication. It has to be stressed, that it has been already questioned, e.g. in [9], if the dBA (A-weighted sound pressure level) measurement of environmental noises is valid for a broad level range. However, it has been commonly accepted and apparent in many publications [5-8], thus used by the authors. Another aspect is the stationarity of noise stimuli used to produce the Lombard effect. The vast majority of environmental noises have varying properties over time and can be highly non-stationary. This is of great importance for preparation of Lombard speech stimuli, as the Lombard effect will vary according to varying noise types causing the effect. An approach taken e.g. in [10] could be followed where the noise used to evoke the Lombard effect is stored in temporal alignment with the recorded Lombard speech, so that it can later be used as a combined stimulus. This approach however is time-consuming and limits the flexibility for future usage of the recordings with other (stationary) noise types. Hence, it was decided to use only one noise type which is stationary enough in order not to produce non-stationary Lombard speech. On the other hand, the usage of highly stationary noise like artificially generated white noise is not ecologically valid. Hence, a babble speech noise has been chosen.

Recording setup

The recording setup block diagram is illustrated on the Figure 1. The recordings took place in the audio studio of the Institute for Media Technology at Technische Universität Ilmenau. The studio is a well acoustically adapted enclosure, where the long term noise level does not exceed $30dBA$. The control room is acoustically separated from the studio.

Playback of the noise stimuli was calibrated beforehand with a dummy head (HEAD acoustics HMS II.3) so that it could be set at which dBA level the noise is presented to the speakers. The same babble speech noise sample was used for calibration as for playback later on. During recordings, the noise stimulus was presented diotically to the speakers at desired dBA level with Beyerdynamic DT 770 closed headphones connected to RME Multiface II sound interface. There is a risk that speakers wearing closed headphones will change their speaking manner. To compensate for this effect, an own-voice feedback to the headphones was implemented. The speakers voice was routed back to the playback system (RME interface) where it was mixed with the noise stimuli. The feedback level was adjusted empirically before actual recordings and in silent conditions so that no difference of own-voice level was perceived when wearing headphones and without them.

Recording of the speech was done by Neumann TLM 103 cardioid microphone, SONY DMX-R100 mixing board and Avid Pro Tools 10 digital audio workstation. The audio was recorded with $44.1kHz$ sampling frequency and $16bit$ linear resolution. The input gain at the mixing board was set for each speaker individually to reach optimal signal to noise ratio. This is due to the natural differences of the voice volume for different speakers. The gain was kept constant till the end of each session to be able to study the voice volume differences for different noise conditions. It has to be stressed that in this setup no absolute levels could be determined for the speakers.

An intercom system was set up to provide a communication channel between a speaker and the person who controlled the recordings. The intercom was switched off during the actual recordings to avoid presentation of unwanted noises to the speakers during the recording phase.

Noise stimuli

The noise sample used during the noise playback calibration and speech recording phase is an artificially created babble speech consisting of multiple overlaid speech recordings. Utterances of sixteen different male and female, French and German speakers were mixed together including pitch-shifted versions of their voices [11].

Text corpus

The text material used during recordings was developed within the EUROM project [12]. In the cited work a set of forty phonemically balanced passages can be found. For the needs of telephony testing the sentence blocks were shortened in a way that each block consists of two to three sentences which corresponds to around 9 s long speech samples. It has to be noted that since shortened versions of the original material were used during recordings, the phonemic balance might be affected. In addition, a furtherly shortened speech set has been provided to serve the listening tests specified in [13]. In this case, the speech samples contain only one sentence each, with the average duration of 2.5 s.

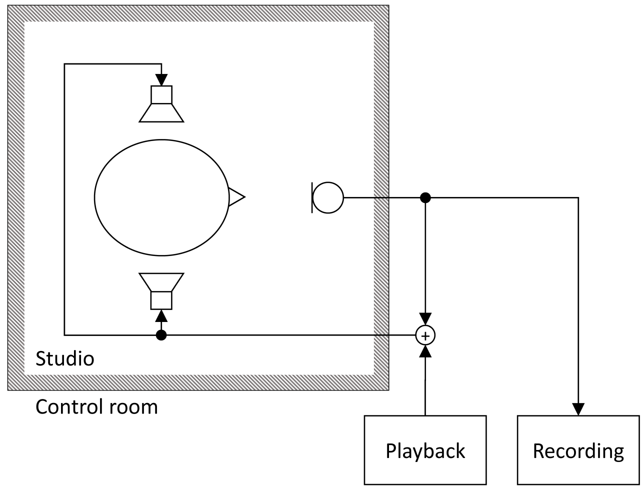

Figure 1: Recording setup

Speakers

A total number of eight native German speakers were invited for the recordings (5 male, 3 female). They were recruited from academic staff and students implying a similar social background. During the recruitment process it was specifically considered that speakers should speak with a dialect-free language, to create a German language database that will be as universal as possible.

Recording procedure

The speakers were invited to the studio for 1.5 h long sessions. Each session consisted of three parts reflecting different noise cases. The order of the cases was always as follows: no noise, 55 dBA and 70 dBA. In the no-noise case the speakers did not wear the headphones. The distance between the speakers mouth and microphone was around $20-30$ cm, and the speakers were asked not to change the distance during the recording phase. They were instructed to read the text clearly but naturally. They were asked to read the prepared sentence blocks one by one with a short break between the blocks. Each block had to be correctly read at least twice so there will be a possibility to choose the better sample for the final dataset. Speakers were allowed to take pauses whenever needed.

Validation of the recordings

To validate the database it was checked if the expected increase in output speech level could be observed. The

Table 1: Level increase in dB for particular speakers in different noise conditions

speaker	$N_1 = 55dBA$	$N_2 = 70dBA$
f1	3.6138	8.4257
f2	4.5866	6.7837
f3	0.7618	4.6977
m1	2.3979	5.2528
m2	0.5733	4.3140
m3	2.8416	7.7097
m4	4.3257	9.0594
m5	3.3372	10.2730

resulting level changes for particular speakers are provided in Table 1. The values were obtained by taking the average level increase for all the cut speech samples for a given speaker for the given noise condition. The standard deviation of the means was usually not higher than 1 dB. The speech levels were determined according to [14].

The reference speech-level increase is modelled and presented e.g. in [5]. According to the equation the level is increased by 3 dB for every 10 dB when the long-term A-weighted noise level exceeds 50 dBA:

$$I(N) = \begin{cases} 0 & \text{for} \quad N < 50\ dBA \\ 0.3(N-50) & \text{for} \quad 50\ dBA \leq N < 77\ dBA \\ 8 & \text{for} \quad N \geq 77\ dBA \end{cases} \tag{1}$$

where:
I - the dB increase in mouth output level due to noise level
N - the long-term A-weighted noise level measured near the user's head position

In the Figure 2 the curves indicating the mean level increase per speaker are depicted. It can be observed that the mean level increase is higher as determined by the presented equation by $1.3dB$ for $N = 55dBA$ and by $1dB$ for $N = 70dBA$. Each speaker was raising the voice level with increasing background noise. This proves the presence of the Lombard effect with regard to changed voice volume. It is worth to note that in a later optional post-processing phase one can equalize the speech levels to the values defined according to the Equation 1. This will ensure that the speech dataset is coherent with regard to speech level, meaning that all speakers are at the same level for given noise conditions. The other voice transformations specific to the Lombard effect will be maintained in that case.

Publication of the speech material

The database has been published via Zenodo and has been assigned a unique Digital Object Identifier (DOI) which should be used for possible citations [16]. Additionally, a GitHub project has been created where information about database updates and related data will be stored [17].

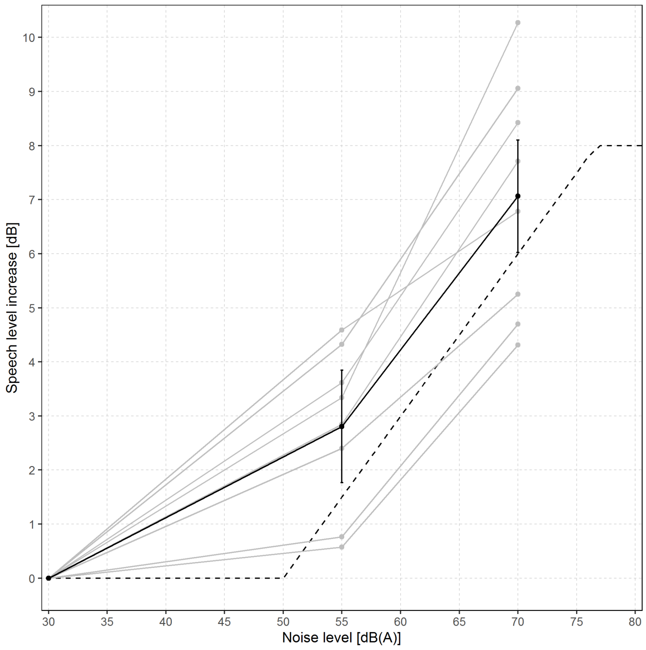

Figure 2: Increase in mouth output level due to noise level. Gray dashed line illustrates the equation taken from ITU-T P.1100. Solid black line depicts mean level increases for all invited speakers (with accompanying 95% confidence intervals). Gray solid lines present level increases for particular speakers.

Applications

Due to the applied headphone presentation, the speech recordings do not include the noise stimuli, and may be mixed with different types of noises in the post-processing phase. Such a material can be used for a variety of perceptual and instrumental tests addressing speech in noisy conditions. These test could address:

- different speech levels
- noise types and levels
- Lombard speech effects
- noise suppression algorithms [13]
- speech recognizers [15]

To reflect the increase of voice volume due to the Lombard effect two approaches can be followed:

1. Apply Equation 1 to different speakers so the speech level is equal for particular noise conditions. In this case it is possible to simulate the conditions with noise levels different than the ones considered in the recording phase.
2. Use the actual level changes which were observed in the recorded material and are provided in Table 1.

Outlook

With the practical experience gained during the recording, it is possible to identify a number of aspects that could be differently addressed in future work. For the recording phase, a randomization of the order of sentence blocks given to the speakers could improve the validity of the dataset in order to account for the fact that speakers tend to change their reading manner over time. Alternatively, a warm up phase could limit this effect so the readers will stabilize their reading style after a few introductory phrases. To limit the unnatural reading style, the speakers can be instructed to imagine that a telephone conversation is taking place so they potentially will try to communicate the content rather than just read it out. An interesting solution to this problem was proposed in [18]. Moreover, it is recommended to extend the validation of the database with some more detailed analysis e.g. by addressing mean value of pitch, phoneme duration and frequency envelope as proposed in [19].

Conclusions

In this paper a Lombard speech dataset has been presented. The invited speakers were native Germans and the text material originates from a popular German speech database [12]. A basic analysis was performed which confirmed the existence of Lombard effect in the speech recordings. The database has been opened to the public and constitutes a valuable dataset which could be utilized in a variety of subjective and instrumental tests.

Acknowledgment

This work was conducted under a BMWi ZIM founded project, STEEM. The project consortium consists of: Technische Universität Ilmenau, HEAD acoustics GmbH and AVM GmbH.

References

[1] Hu, Y., Loizou, P. C.: Subjective comparison and evaluation of speech enhancement algorithms, Speech Communication, 2007, 49, 588-601

[2] Li, J., Yang, L., Zhang, J., Yan, Y.: Comparative intelligibility investigation of single-channel noisereduction algorithms for Chinese, Japanese, and English, The Journal of the Acoustical Society of America, 2011, 129, 3291-3301

[3] Lombard, É.: Le signe de l'elevation de la voix Annals maladiers oreille, Larynx, Nez, Pharynx, 1911, 37, 101 – 119

[4] Junqua, J.-C.: The influence of acoustics on speech production: A noise-induced stress phenomenon known as the Lombard reflex, Speech Communication, 1996, 20, 13-22

[5] ITU-T Rec. P.1100 - Narrow-band hands-free communication in motor vehicles, International Telecommunication Union, 2015

[6] ETSI EG 202 396-1: Background noise simulation technique and background noise database European Telecommunications Standards Institute, 2008

[7] ETSI TS 103 224: A sound field reproduction method for terminal testing including a background noise database, European Telecommunications Standards Institute, 2014

[8] NIOSH: Occupational Noise Exposure, The National Institute for Occupational Safety and Health, Cincinnati, USA, 1998

[9] Richard L. St. Pierre, J., Maguire, D. J.: The impact of A-weighting sound pressure level measurements during the evaluation of noise exposure, Proc. of the Twentieth National Conference on Noise Control Engineering, Noise-Con 2004

[10] Ullmann, R., Bourlard, H., Berger, J., Llagostera Casanovas, A.: Noise Intrusiveness Factors in Speech Telecommunications, Proc. of the AIA-DAGA 2013 International Conference on Acoustics, 2013, 436-439

[11] Raake A., Katz B.: Measurement and Prediction of Speech Intelligibility in a Virtual Chat Room, Proc. 2nd ISCA/DEGA Tutorial and Research Workshop on Perceptual Quality of Systems, 2006

[12] EUROM project `http://www.phon.ucl.ac.uk/shop/eurom1.php`

[13] ITU-T Rec. P.835 - Subjective test methodology for evaluating speech communication systems that include noise suppression algorithm, International Telecommunication Union, 2003

[14] ITU-T Rec. P.56 - Objective measurement of active speech level International Telecommunication Union, 2011

[15] Junqua, J.-C.: The Lombard reflex and its role on human listeners and automatic speech recognizers, Journal of the Acoustical Society of America, 1993, 1, 510-524

[16] Zenodo DOI: 10.5281/zenodo.48713 `https://zenodo.org/record/48713`

[17] Lombard speech database for German language GitHub project, URL: `https://github.com/Telecommunication-Telemedia-Assessment/Lombard-Speech-database.git`

[18] Boril, H., Boril, T., Pollák, P.: Methodology of Lombard speech database acquisition: Experiences with CLSD, Proc. of the Fifth Conference on Language Resources and Evaluation – LREC'06, 2006

[19] Vlaj, D., Markuš, A. Z., Kos, M., Kačič, Z.: Acquisition and Annotation of Slovenian Lombard Speech Database, Proc. of the Seventh International Conference on Language Resources and Evaluation - LREC'10, 2010