Interspeech 2024
1-5 September 2024, Kos, Greece

Does the Lombard Effect Matter in Speech Separation? Introducing the Lombard-GRID-2mix Dataset

Iva Ewert[1], *Marvin Borsdorf*[1], *Haizhou Li*[3,1,2], *Tanja Schultz*[4]

[1]Machine Listening Lab (MLL), University of Bremen, Germany
[2]Department of Electrical and Computer Engineering, National University of Singapore, Singapore
[3]SDS, SRIBD, The Chinese University of Hong Kong, Shenzhen, China
[4]Cognitive Systems Lab (CSL), University of Bremen, Germany
iewert@uni-bremen.de, marvin.borsdorf@uni-bremen.de

Abstract

Inspired by the human ability of selective listening, speech separation aims to equip machines with the capability to disentangle cocktail party soundscapes into the individual sound sources. Recently, neural network based algorithms have been studied to work reliably under various conditions. However, to the best of our knowledge, a change in the speaking style has not yet been studied. The Lombard effect, a reflexive change in speaking style triggered by noisy environments, is a typical behavior in everyday conversational situations. In this work, we introduce a new first of its kind dataset, called Lombard-GRID-2mix, to study speech separation for two-speaker mixtures on normal speech and Lombard speech. In a comprehensive study, we show that speech separation systems can be equipped to work for both normal speech and Lombard speech. We apply a carefully designed finetuning method to enable the system to work even if noise is present in the Lombard speech for different SNR ratios.

Index Terms: Speech separation, cocktail party problem, selective auditory attention, Lombard effect, noisy speech

1. Introduction

Humans have the ability to almost effortlessly focus on a particular sound source within a given mixture signal, referred to as selective auditory attention. Cherry et al. [1] formulated this in 1953 by introducing the cocktail party problem. Equipping machines with this ability, however, still represents an unsolved research problem. Solutions to the problem are motivated by two main strands: (a) the application in smart hearing aids (b) its potential as a front-end for downstream tasks such as automatic speech recognition (ASR) and speaker verification. The research that aims to develop machines that mimic this human ability is called speech separation. The goal of speech separation is to disentangle a given mixture signal into the individual speech signals. For this, two main algorithms have been introduced: Blind source separation (BSS) [2, 3, 4, 5, 6, 7] and target speaker extraction (TSE) [8, 9, 10]. BSS separates a given mixture signal of overlapping sources into all individual sources in one step. This approach usually requires the number of contributing sources to be known in advance. Furthermore, the assignment of output channels to the sound sources is ambiguous. This so called permutation problem can be solved by applying the permutation invariant training (PIT) [3, 4] method. In contrast to BSS, TSE extracts solely the voice of a particular target speaker from a given mixture signal. To perform this method, a supplementary information commonly in form of a target speaker's speech sample is used.

Initially, speech separation has been studied in the frequency-domain in which the algorithms operate on the Short-Time Fourier Transform representation of the mixture signal [2, 3, 11]. In recent time, end-to-end approaches operating in the time-domain have been introduced [5, 6, 7]. The latter elegantly circumvent the inaccuracy of using the phase information of the mixture signal for the reconstruction of the separated sources in the time-domain and, in addition, allow to directly process the raw audio signal.

In everyday conversational situations, speech separation systems have to handle typical changes in the conversation and soundscape, such as unknown voices, different numbers of overlapping voices, and varying background noise. Another typical behavior of human speakers is a change in the speaking style to enhance the intelligibility of their speech in adverse conditions. While speakers who communicate with hearing-impaired listeners adapt a so-called clear speaking style characterized by overly articulated speech, speakers who are faced by noisy environments tend to increase their vocal effort to overcome the communication barrier [12], which is called the Lombard effect [13]. Typically, the effect is characterized by a rise in the sound volume, an alteration of the fundamental frequency, an increase in the vowel duration, and a change in the formant frequencies of the speech signals [14, 15, 16]. Usually, as the acquisition of speech in noisy environments requires a complex recording setup, noise is, if at all, artificially added to normal speech in the datasets to develop noise-robust speech separation systems [17, 18]. In real-life scenarios, however, speech separation systems have to cope with the combination of noise and Lombard speech, as this particular speaking style adaptation is usually accompanied by noisy environments.

In the field of ASR, recent studies [14, 19, 20, 21] investigated the effect of Lombard speech on selected ASR systems. Ma et al. [20] investigated the impact of the Lombard effect on audio-only, video-only, and audio-visual speech recognition and demonstrated the benefits of integrating Lombard speech into the system development. The influence of the Lombard effect has also recently attracted attention in the field of speech enhancement [22, 23, 24]. Speech enhancement aims to improve the quality of speech signals that are adversely affected for example by noise. In Michelsanti et al. [23], a performance gap of approx. 5 dB between systems trained on normal speech and systems trained on Lombard speech was identified, if tested on Lombard speech. Consequently, the existence of similar phenomena in the field of speech separation can be assumed. However, to the best of our knowledge, studies on this topic have not

10.21437/Interspeech.2024-1131

yet been carried out.

In this work, we study the influence of the Lombard effect on speech separation systems. For this, we propose and construct a new first of its kind dataset, called Lombard-GRID-2mix. The dataset consists of two-speaker mixtures generated on the basis of speech material derived from the Audio-Visual Lombard GRID Corpus [25]. Lombard-GRID-2mix is designed to match the BSS paradigm and is divided into two sets containing mixtures of normal or Lombard speech, respectively. On the basis of this dataset, we develop two systems and evaluate each on both normal and Lombard speech. In a comprehensive experimental setup, we investigate the benefits of integrating Lombard speech in the system development and apply different finetuning strategies. We record a performance degradation of the system trained on normal data in the presence of Lombard speech and identify a suitable finetuning strategy to overcome this problem. As Lombard speech is triggered by the occurrence of noise in real world conversational situations, we extend our dataset with four additional versions of the Lombard subset that integrate speech-shaped noise (SSN) at different signal-to-noise-ratios (SNRs). We find that the performances of all systems drop drastically on Lombard speech in noise, but we can substantially mitigate this degradation by integrating Lombard speech into a carefully designed training process.

The rest of the paper is organized as follows. In Section 2, we introduce the Lombard-GRID-2mix dataset and explain its construction. In Section 3, we describe the experimental setup, followed by Section 4, which presents and discusses the results. Section 5 concludes our study and highlights some future work. We make all scripts to simulate the Lombard-GRID-2mix dataset publicly available[1].

2. Construction of Lombard-GRID-2mix

To compare the speech separation performance on normal speech versus Lombard speech, an appropriate dataset is required. To achieve this, we propose a new first of its kind dataset, called Lombard-GRID-2mix, which provides two-speaker speech mixtures and the respective single speaker voices in both normal speech and Lombard speech. The dataset is derived from the Audio-Visual Lombard GRID Corpus [25].

2.1. The Audio-Visual Lombard GRID Speech Corpus

The Audio-Visual Lombard GRID Speech Corpus [25] contains recorded utterances produced by 54 (30 female and 24 male) native speakers of British English between the age of 18 and 30. The construction of the recorded sentences followed the Grid corpus syntax [26]. From a set of 30,000 sentences, a subset of 50 unique sentences was assigned to a speaker and recorded in normal (N) and in Lombard (L) speech. To trigger the Lombard speech, SSN was presented to the speaker at a sound pressure level of 80 dB via headphones. The speaker read the sentences to a human listener to ensure the intelligibility of the speech, which represented one of the prerequisites for the Lombard effect. Every five to seven sentences, the human listener asked the speaker to repeat the sentence while pretending not to understand the last utterance, which represented the first form of feedback. In addition, the produced speech of the speaker was presented to the speaker at a carefully adjusted level, as speakers regulate their speaking style also by considering the perceived level of their own speech.

[1] https://github.com/iewert/LombardGRID2mix

2.2. Design of subsets and two-speaker mixtures

The newly proposed Lombard-GRID-2mix dataset consists of two sets: Lombard-GRID-2mix-Normal and Lombard-GRID-2mix-Lombard, each containing two-speaker mixtures and their corresponding single sources in the respective speaking style. Each set provides training, validation, and test subsets. The Audio-Visual Lombard GRID Speech Corpus is divided into (training/validation) and test speakers by applying a 80/20 % split. The final test set comprises all 50 utterances from each test speaker. The pool of utterances for the training and validation speakers is divided by a 75/25 % split per speaker to obtain the training and validation set. The utterances in the test subset are different to the (training/validation) subset. The gender distribution of the Audio-Visual Lombard GRID Speech Corpus is maintained in all subsets.

Since the combination of all speakers and utterances to simulate the two-speaker mixtures would lead to high computational costs, we apply the following strategy: First, we introduce different groups. Each group consists of five utterances per speaker in the considered set. Second, within these groups, we combine all utterances of different speakers. Third, for each utterance combination, a SNR is randomly sampled. The SNR value is drawn from a uniform distribution in the range of $[0, 5]$ dB. Fourth, the fore- and background speakers are randomly selected by assigning $\pm\frac{\mathrm{SNR}}{2}$. This strategy ensures a zero mean and a signal-level difference between both speakers based on the sampled SNR value. Fifth, the scripts of Isik et al. [27], originally implemented for the creation of the wsj0-2mix and wsj0-3mix datasets [2], are used to simulate the actual mixture data. For the presented experiments, a sampling frequency of 8 kHz and the "min" mode are applied. In the "min" mode, the longer utterance is cut to the length of the shorter utterance when mixing two utterances, resulting in highly overlapped speech.

The final Lombard-GRID-2mix dataset comprises 109.5, 29, and 9.2 hours (normal) and 118.3, 31.2, and 9.9 hours (Lombard) of total audio data for training, validation, and testing, respectively. The total amount of audio data in the Lombard set is higher due to an increased utterance duration because of the adapted speaking style.

2.3. Addition of speech-shaped noise

As Lombard speech is typically accompanied by noise, we repeat the simulation of the Lombard subset (Lombard-GRID-2mix-Lombard) with additional background noise at four different SNR ranges. For this, we simulate speech-shaped noise (SSN) that matches the Lombard recording condition of the Audio-Visual Lombard GRID speech corpus. SSN is a stationary noise type that follows the spectrum of human speech, obtained by applying the Discrete Fast Fourier Transform. To ensure the independence of the noise signals from the speech material given in the mixtures, we utilize a different dataset for noise generation, namely the Wall Street Journal (WSJ0) [28] corpus. For each speaker in the WSJ0 corpus, a noise signal is created by extracting the long-term spectrum of the respective speech material, followed by phase randomization. A SSN sample is assigned to each two-speaker mixture signal in the Lombard-GRID-2mix-Lombard set. To simulate the two-speaker mixture signal including SSN in the background, the three-speaker mixture simulation scripts from [27] are used in which the SSN is considered as a third speaker's signal. In each of the four

Table 1: *Speech separation results for different systems and training methods. All finetuned systems are pre-trained on the Lombard-GRID-2mix-Normal subset. Systems #5 and #6 are finetuned with a learning rate of $1e^{-5}$ and $1e^{-4}$, respectively. The results are reported in terms of SI-SDRi (dB) as mean over all samples in the normal speech and Lombard speech test sets, respectively. In addition, we report the mean performance over both test sets. The higher the SI-SDRi, the better.*

Number	System	Training method	Trained on		Results on test sets		
			Lombard-GRID-2mix-		Lombard-GRID-2mix-		
			Normal	Lombard	Normal	Lombard	Mean
1	Conv-TasNet-N	from scratch	✓	✗	14.71	12.25	13.48
2	Conv-TasNet-L	from scratch	✗	✓	**15.05**	**15.02**	**15.03**
3	DPRNN-N	from scratch	✓	✗	14.70	12.97	13.84
4	DPRNN-L	from scratch	✗	✓	14.19	14.45	14.32
5	DPRNN-FT-I	finetuning (from #3)	✗	✓	14.49	14.04	14.26
6	DPRNN-FT-II	finetuning (from #3)	✓	✓	**15.04**	**14.21**	**14.62**

versions of the noisy Lombard subset, the SNR between both speakers is kept the same as in the Lombard subset without noise. Each noisy version is characterized by a deterministic noise level from $\{3, -2.5, -8, -13.5\}$ dB at which the SSN is additionally mixed in. The noise levels were selected on the basis of typical noises from electrical devices at a distance of 1 m from a human conversation at average conversation volume. In the resulting mixture, the SNR between one of the two speakers and the noise lies in the following non-overlapping ranges: $[-5.5, -0.5]$, $[0, 5]$, $[5.5, 10.5]$, and $[11, 16]$ dB.

3. Experimental setup

In our study, we develop and evaluate nine speech separation systems on the Lombard-GRID-2mix dataset. We train the systems from scratch and apply finetuning methods. The system performance of the systems is evaluated on different test sets. The experiments are implemented in Python using PyTorch.

3.1. Network architectures

The systems in our experiments are based on the widely studied Conv-TasNet [5] and DPRNN [6] architectures. Both architectures work in the time-domain and apply the mask estimation approach as follows: The encoder transforms the input mixture signal into a latent representation. The separation part of the system estimates masks for each of the sources of the input mixture, which are subsequently element-wise multiplied with the latent representation of the input mixture to obtain the separated signals. The decoder reconstructs the waveform of the separated sources.

The systems are implemented using the audio source separation toolkit Asteroid[2] [29]. The configuration of the Conv-TasNet is chosen according to the best system identified by Luo et al. [5]. Due to limited computation power, the configuration of the DPRNN with a chunk size of 100 and a kernel length of 16 is chosen. The used implementations of the Conv-TasNet and the DPRNN comprise 5.05 million and 3.65 million parameters, respectively.

3.2. Training and evaluation procedure

The speech separation systems are trained from scratch or finetuned on different subsets of the Lombard-GRID-2mix dataset in a supervised learning fashion. During training and finetuning, the negative value of the scale-invariant source-to-distortion ratio (SI-SDR) [30] (Eq. 1) is used as the loss function:

$$SI\text{-}SDR \approx 20 \cdot \log_{10} \left(\frac{\left\| \frac{\hat{s}^T \cdot s}{\|s\|^2 + \varepsilon} \cdot s \right\|}{\left\| \frac{\hat{s}^T \cdot s}{\|s\|^2 + \varepsilon} \cdot s - \hat{s} \right\| + \varepsilon} + \varepsilon \right) \quad (1)$$

with s being the ground truth signal and $\hat{s}$ denoting the reconstructed signal. We add $\varepsilon = 1e^{-8}$ to stabilize the loss function. To circumvent the source permutation problem, utterance-level PIT [3] is applied. The training routine is implemented based on a publicly available framework[3].

The data is split into chunks of 4 s length, smaller samples between two and four seconds are padded with zeros while even smaller samples are discarded. The chunks are merged into mini-batches. The Conv-TasNet and the DPRNN based system are trained with a batch size of 32 and 64, respectively. Adam [31] is used as optimizer with an initial learning rate of $1e^{-3}$ and a weight decay of $1e^{-5}$. For the finetuning experiments, the initial learning rate is set to $1e^{-5}$ or $1e^{-4}$ depending on whether Lombard speech is used solely or altogether with normal speech data. The learning rate is halved after two subsequent epochs without any improvement on the validation data. The minimum learning rate is set to $1e^{-8}$. After six epochs with no improvement on the validation loss, the training is stopped. The maximum number of epochs is set to 150 for both strategies, training from scratch and finetuning.

All systems are evaluated on all test sets. During evaluation, the utterances are fed into the systems with their entire length. The estimated separated signals are compared with their respective ground truth signals by calculating the SI-SDR [30]. The results are reported in terms of SI-SDR improvement (SI-SDRi) which is calculated as the difference between the SI-SDR of the estimated signals and the SI-SDR of the input mixture. SI-SDRi results are stated as the mean over all utterances in the considered test set.

4. Results and discussion

In total, we trained nine speech separation systems. While two systems were trained on Lombard-GRID-2mix-Normal only, two other systems were trained solely on Lombard-GRID-2mix-Lombard. The other systems were developed by applying different finetuning methods to one of the pretrained systems. The systems are evaluated on Lombard-GRID-2mix-Normal and Lombard-GRID-2mix-Lombard test sets (see Table 1).

[2] https://github.com/asteroid-team/asteroid

[3] https://github.com/funcwj/conv-tasnet/

Table 2: *Speech separation results for different systems and training methods on noisy Lombard speech (Lombard-noisy). All finetuned systems are pre-trained on the Lombard-GRID-2mix-Normal subset. While system #5 is finetuned with a learning rate of* $1e^{-5}$*, systems #6-9 are finetuned with a learning rate of* $1e^{-4}$*. The results are reported in terms of SI-SDRi in dB as mean over all samples in the respective Lombard-noisy test set. The SNR range (dB) between the speakers and the noise given in the mixtures is indicated by* $N^{[\mathit{Higher\ limit}}_{\mathit{Lower\ limit}]}$*. In addition, we report the mean performance over all Lombard-noisy test sets. The higher the SI-SDRi, the better.*

Number	System	Training method	Trained on Lombard-GRID-2mix-				Results on test sets				
			Normal	Lombard	Lombard-noisy		Lombard-noisy				
					$N^{[-0.5]}_{[-5.5]}$	$N^{[16]}_{[11]}$	$N^{[-0.5]}_{[-5.5]}$	$N^{[5]}_{[0]}$	$N^{[10.5]}_{[5.5]}$	$N^{[16]}_{[11]}$	Mean
3	DPRNN-N	from scratch	✓	✗	✗	✗	-0.82	0.15	**3.50**	**8.31**	**2.78**
4	DPRNN-L	from scratch	✗	✓	✗	✗	**-0.30**	**0.42**	1.75	8.12	2.39
5	DPRNN-FT-I	finetuning (from #3)	✗	✓	✗	✗	-0.63	0.22	3.34	8.14	2.77
6	DPRNN-FT-II	finetuning (from #3)	✓	✓	✗	✗	-0.62	0.37	3.69	8.56	3.00
7	DPRNN-FT-III	finetuning (from #3)	✓	✗	✗	✓	0.23	2.93	8.10	**11.67**	5.73
8	DPRNN-FT-IV	finetuning (from #3)	✓	✗	✓	✗	4.32	**8.12**	6.91	9.62	7.24
9	DPRNN-FT-V	finetuning (from #3)	✓	✗	✓	✓	**4.44**	6.27	**8.95**	11.34	**7.75**

Systems #1 and #3 that were trained on normal speech data only show a performance drop of 2.46 dB and 1.73 dB if tested on Lombard speech. Here, the Conv-TasNet shows a lower generalizability to the unseen speaking style. Training the systems purely on Lombard speech (sytems #2 and #4) leads to a higher performance on the Lombard test set by 2.77 dB and 1.48 dB. Surprisingly, the performance on normal speech test data does not degrade substantially and even improves in the case of the Conv-TasNet (system #2) compared to the systems trained on normal speech data (systems #1 and #3). Considering the mean performance over both test sets, system #2 shows the best speech separation results.

Table 1 shows the results for two different finetuning methods: (i) finetuning the system solely on Lombard data (system #5 with an initial learning rate of $1e^{-5}$); (ii) finetuning the system on a balanced mixture of normal speech data and Lombard speech data (system #6 with an initial learning rate of $1e^{-4}$). Due to computational limitations and the considerably lower training time of the DPRNN, the finetuning experiments were conducted on the DPRNN system only. Both finetuning experiments were carried out by using the pre-trained system #3. System #6 outperforms system #5 that lacks training information of normal data, under both test conditions. Compared to its basis (system #3), system #6 shows a strong performance improvement on the Lombard speech data and a slight increase in performance on the normal speech data. Based on these results, finetuning a speech separation system that was trained on normal speech data only with additional Lombard speech data represents a promising approach to increase its robustness against the occurrence of the Lombard effect.

Nevertheless, the previously discussed results did not take into account the presence of noise, which is typically given in real-world Lombard speech scenarios. Table 2 shows the speech separation results for Lombard speech mixtures that include background noise with different SNR levels, processed by a DPRNN system. Overall, the performance of all systems drops substantially in the presence of noise. This degradation intensifies if the noise level increases. Remarkably, system #3 outperforms system #4 under two conditions, although it was trained on normal speech only. The finetuned system #6 shows a slightly increased performance on all noisy test sets compared to the basis system #3. However, systems #3-6 lack robustness against the occurrence of noise.

To counteract this limitation, we incorporated Lombard speech data that contained noise into the finetuning. For this, we used the lowest and highest SNR ranges, i.e., [-5.5, -0.5] and [11, 16] dB, respectively. Both systems #7-8 were simultaneously trained on normal speech and Lombard speech (including noise), as this strategy has proved to be superior in the previous experiments. Both systems show an increased performance on all noisy test sets compared to the previous systems #3-6. Comparing them among each other, system #7 that was trained on less prominent noise outperforms the system #8 on the two test sets with the lowest noise levels. Conversely, system #8 shows superior performance on the two test sets with stronger noise levels.

To combine the strengths of both systems, system #9 was finetuned on normal speech and Lombard speech that contains both noise levels while maintaining the overall training data size equal to the training of systems #7-8. The results indicate that system #9 outperforms the other systems on two noise levels as well as on the mean performance across all four noisy test sets.

5. Conclusion and future work

To the best of our knowledge, we are the first who studied the influence of the Lombard effect on speech separation systems. We proposed a new first of its kind dataset, called Lombard-GRID-2mix, to simulate two-speaker cocktail party mixtures given in both normal speech and Lombard speech. To better reflect real-world Lombard speech situations, we extend our dataset to include Lombard speech and noise in four different SNR ratios. In total, we developed nine different speech separation systems and show that a system can be equipped to work for both normal speech and Lombard speech. With a carefully designed training procedure, systems show better performance for Lombard speech in the presence of noise and can work under different noise levels.

In our future work, we will investigate new strategies to further improve the robustness of speech separation systems in the presence of Lombard speech and noise. In particular, we would like to focus on scenarios with variable noise types, noise ranges, and reverberation. Moreover, we aim to extend this analysis to a broader variety of speech separation systems that use recent techniques, such as attention.

6. Acknowledgements

This research is funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy (University Allowance, EXC 2077, University of Bremen).

7. References

[1] E. C. Cherry, "Some experiments on the recognition of speech, with one and with two ears," *The Journal of the Acoustical Society of America*, vol. 25, no. 5, pp. 975–979, 1953.

[2] J. R. Hershey, Z. Chen, J. L. Roux, and S. Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in *ICASSP*, 2016, pp. 31–35.

[3] M. Kolbæk, D. Yu, Z.-H. Tan, and J. Jensen, "Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 10, pp. 1901–1913, 2017.

[4] D. Yu, M. Kolbæk, Z.-H. Tan, and J. Jensen, "Permutation invariant training of deep models for speaker-independent multi-talker speech separation," in *ICASSP*, 2017, pp. 241–245.

[5] Y. Luo and N. Mesgarani, "Conv-TasNet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 8, pp. 1256–1266, 2019.

[6] Y. Luo, Z. Chen, and T. Yoshioka, "Dual-path RNN: Efficient long sequence modeling for time-domain single-channel speech separation," in *ICASSP*, 2020, pp. 46–50.

[7] C. Subakan, M. Ravanelli, S. Cornell, M. Bronzi, and J. Zhong, "Attention is all you need in speech separation," in *ICASSP*, 2021, pp. 21–25.

[8] M. Delcroix, K. Zmolikova, T. Ochiai, K. Kinoshita, S. Araki, and T. Nakatani, "Compact network for speakerbeam target speaker extraction," in *ICASSP*, 2019, pp. 6965–6969.

[9] M. Ge, C. Xu, L. Wang, E. S. Chng, J. Dang, and H. Li, "SpEx+: A complete time domain speaker extraction network," in *INTERSPEECH*, 2020, pp. 1406–1410.

[10] J. Zhang, C. Zorilă, R. Doddipatla, and J. Barker, "Time-domain speech extraction with spatial information and multi speaker conditioning mechanism," in *ICASSP*, 2021, pp. 6084–6088.

[11] C. Xu, W. Rao, X. Xiao, E. S. Chng, and H. Li, "Single channel speech separation with constrained utterance level permutation invariant training using grid LSTM," in *ICASSP*, 2018, pp. 6–10.

[12] E. Godoy, M. Koutsogiannaki, and Y. Stylianou, "Approaching speech intelligibility enhancement with inspiration from lombard and clear speaking styles," *Computer Speech & Language*, vol. 28, no. 2, pp. 629–647, 2014.

[13] E. Lombard, "Le signe de l'élévation de la voix (translated from french)," *Ann. des Mal. l'oreille du larynx*, vol. 37, no. 2, pp. 101–119, 1911.

[14] J.-C. Junqua, "The Lombard reflex and its role on human listeners and automatic speech recognizers," *The Journal of the Acoustical Society of America*, vol. 93, no. 1, pp. 510–524, 1993.

[15] W. V. Summers, D. B. Pisoni, R. H. Bernacki, R. I. Pedlow, and M. A. Stokes, "Effects of noise on speech production: Acoustic and perceptual analyses," *The Journal of the Acoustical Society of America*, vol. 84, no. 3, pp. 917–928, 1988.

[16] Y. Lu and M. Cooke, "Speech production modifications produced by competing talkers, babble, and stationary noise," *The Journal of the Acoustical Society of America*, vol. 124, no. 5, pp. 3261–3275, 2008.

[17] G. Wichern, J. Antognini, M. Flynn, L. R. Zhu, E. McQuinn, D. Crow, E. Manilow, and J. L. Roux, "WHAM!: Extending speech separation to noisy environments," in *INTERSPEECH*, 2019, pp. 1368–1372.

[18] J. Cosentino, M. Pariente, S. Cornell, A. Deleforge, and E. Vincent, "Librimix: An open-source dataset for generalizable speech separation," *arXiv preprint arXiv:2005.11262v1*, 2020.

[19] R. Marxer, J. Barker, N. Alghamdi, and S. Maddock, "The impact of the Lombard effect on audio and visual speech recognition systems," *Speech Communication*, vol. 100, pp. 58–68, 2018.

[20] P. Ma, S. Petridis, and M. Pantic, "Investigating the Lombard effect influence on end-to-end audio-visual speech recognition," in *INTERSPEECH*, 2019, pp. 4090–4094.

[21] S. U. Maheswari, A. Shahina, and A. N. Khan, "Understanding Lombard speech: a review of compensation techniques towards improving speech based recognition systems," *Artificial Intelligence Review*, vol. 54, no. 4, pp. 2495–2523, 2020.

[22] D. Michelsanti, Z.-H. Tan, S. Sigurdsson, and J. Jensen, "Deep-learning-based audio-visual speech enhancement in presence of Lombard effect," *Speech Communication*, vol. 115, pp. 38–50, dec 2019.

[23] D. Michelsanti, Z.-H. Tan, S. Sigurdsson, and J. Jensen, "Effects of Lombard reflex on the performance of deep-learning-based audio-visual speech enhancement systems," in *ICASSP*, 2019, pp. 6615–6619.

[24] S. A. Nossier, J. Wall, M. Moniri, C. Glackin, and N. Cannings, "An experimental analysis of deep learning architectures for supervised speech enhancement," *Electronics*, vol. 10, no. 1, p. 17, 2020.

[25] N. Alghamdi, S. Maddock, R. Marxer, J. Barker, and G. J. Brown, "A corpus of audio-visual Lombard speech with frontal and profile views," *The Journal of the Acoustical Society of America*, vol. 143, no. 6, pp. EL523–EL529, 2018.

[26] M. Cooke, J. Barker, S. Cunningham, and X. Shao, "An audio-visual corpus for speech perception and automatic speech recognition," *The Journal of the Acoustical Society of America*, vol. 120, no. 5, pp. 2421–2424, 2006.

[27] Y. Isik, J. Roux, S. Chen, and J. Hershey, "Scripts to create wsj0-2 speaker mixtures," MERL Research, retrieved October 23, 2023. [Online]. Available: https://www.merl.com/demos/deepclustering/create-speaker-mixtures.zip

[28] J. Garofolo, D. Graff, D. Paul, and D. Pallett, "Csr-i (wsj0) complete ldc93s6a," *Web Download. Philadelphia: Linguistic Data Consortium*, vol. 83, 1993.

[29] M. Pariente, S. Cornell, J. Cosentino, S. Sivasankaran, E. Tzinis, J. Heitkaemper, M. Olvera, F.-R. Stöter, M. Hu, J. M. Martín-Doñas, D. Ditter, A. Frank, A. Deleforge, and E. Vincent, "Asteroid: The PyTorch-based audio source separation toolkit for researchers," in *INTERSPEECH*, 2020, pp. 2637–2641.

[30] J. L. Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "SDR – half-baked or well done?" in *ICASSP*, 2019, pp. 626–630.

[31] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *ICLR*, 2015.