Audio, Visual and Audiovisual intelligibility of vowels produced in noise

Maëva Garnier

Univ. Grenoble Alpes, CNRS, Grenoble INP, GIPSA-lab, 38000 Grenoble, France

maeva.garnier@gipsa-lab.fr

Abstract

Why do speakers amplify articulatory movements when communicating in noisy environments? This study examines the hypothesis that hyper-articulation contributes to improved vowel intelligibility in audio, visual and audiovisual domains. A perceptual test was conducted with Audio-Only (AO), Visual-Only (VO) and Audiovisual (AV) stimuli of vowels produced in conversational and Lombard speech. The average score of vowel recognition was significantly increased in Lombard speech, compared to normal speech, for all perceptual modalities (AO, VO and AV). Specifically, the distinctive features of vowel height and backness were better perceived in Lombard speech in both the audio and visual domains. Changes in speech articulation in noise did not affect the perception of the rounding feature in the visual domain, but degraded it in the audio domain. On the contrary, the perception of the spreading feature was decreased in Lombard speech in the visual domain, but improved in the audio domain.

Index Terms: Lombard speech, hyper-articulation, multimodality, audiovisual intelligibility, perception

1. Introduction

Speech produced in a noisy environment, also known as Lombard speech, exhibits several modifications compared to speech produced in quiet. In particular, speakers speak louder and at a higher pitch in noise. The spectrum of their voice is shifted towards medium frequencies, and the first formant of vowels is increased [1, 2, 3, 4, 5]. More recent studies also showed how Lombard speech is characterized by increased amplitude and velocity of articulatory movements of the jaw, lips and tongue [6, 7, 8, 9, 10].

Perceptual studies have shown that all these acoustic and articulatory changes have a positive impact on speech audiovisual intelligibility [11, 7]. A significant perceptual benefit of Lombard speech, compared to conversation speech produced in quiet, has already been found in the audio domain alone, for the comprehension of words and sentences [1, 12, 13, 14, 4, 6] and more specifically, of vowels and voiced consonants [2]. Furthermore, comparison of this perceptual benefit in audiovisual and purely auditory conditions showed that visible articulatory changes in Lombard speech also contribute, in most cases, to its improved intelligibility in the visual domain (three participants in [15]; hard listening condition in [11]).

However, all these studies focused on words or sentences perception and examined the global consequences of the Lombard effect on speech intelligibility. Yet, little is known about the more specific consequences of the Lombard effect on segment intelligibility. In production, several studies have already documented, for several languages, the acoustic modifications of Lombard speech at the phonemic level [3, 1, 2, 16]. In particular, they showed that vowel systems are not only shifted to higher F1 frequencies, but that they undergo a more global reorganization, affecting the acoustic distance between the different vowel categories – and thus, potentially their perceptual contrast. Similarly, segment-specific modifications of lip rounding, spreading, protrusion or compression have been reported in previous studies of Lombard speech [17, 10]. The goal of this study is therefore to examine in detail the consequences that the Lombard effect can have on the perceptual recognition and discrimination of vowels, in the audio, visual and audiovisual domains.

2. Material and Method

2.1. Audiovisual recordings

Three native French-speaking women were recorded while playing an interactive game in silence and 85 dB broadband noise [5]. The game consisted of communicating information to the experimenter, who stood 2.5 m in front of the speaker. It required to use 15 highly confusable words ([lala], [lela], [lale], [lali], [lila], [lyla], [laly], [lalu], [lula], [pala], [lapa], [bala], [laba], [mala], and [lama]), in unpredictable utterances, so that the recorded speech was produced with a search for intelligibility. The audio signal was recorded with an AKG microphone placed 20cm away from the lips, then digitized at a rate of 44.1kHz. Front videos of the speaker, focused on the lips, were synchronously recorded at a rate of 25 images/s.

2.2. Stimuli

The five target words [lala], [lela], [lila], [lyla] and [lula] produced in initial and final position of the utterances were selected to constitute the stimuli of a perceptual test. These words were segmented from the video recording with their preceding determiner “La”. Thus, the five vowels /a/, /e/, /i/, /y/ and /u/ of interest were situated in the central syllable of a same syntagm (“La l_la”). This way, 120 audiovisual stimuli were created (5 words * 2 positions within the utterance * 2 conditions * 2 repetitions of each condition * 3 speakers). All these stimuli were normalized in intensity, according to their average sound pressure level, over the whole phrase (“La l_la”). Thus, the stimuli of Lombard speech (clean of any ambient noise in which it was produced) were as loud as the stimuli of speech produced in silence. In a second step, both the normal and Lombard stimuli were degraded by the same broadband noise used in the recording session, with a signal to noise ratio of -15 dB (based on [18]). Two additional sets of 120 Audio Only (AO) and 120 Visual Only (VO) stimuli were created from these original AV stimuli. For AO stimuli, the video stream was replaced by a

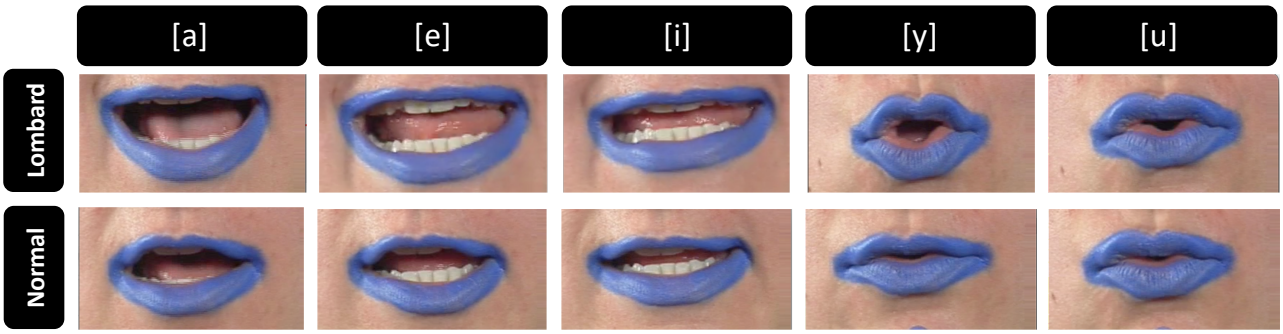

Figure 1: *Examples of the 5 French vowels [a], [e], [i], [y] and [u] produced by the same speaker in conversational and Lombard speech.*

black screen. For VO stimuli, the audio stream was replaced with silence.

2.3. Perceptual test

29 listeners (21 males and 8 females, of 38 ± 13 years old), participated in the perceptual test. All were native French speakers and had no hearing or vision problems. Participants were seated in a quiet room, 30cm away from the screen. The videos were presented with the software Presentation (Neurobehavioral systems) in a 1:1 ratio with reality, so that the conditions of perception were as close as possible to a real face-to-face interaction situation. Participants listened to the stimuli with headphones (Sennheiser HD 250 linear II). The presentation level of the stimuli was calibrated before the experiment to a comfortable level of 75dB, using an artificial ear (Bruël&Kjaer 4153). The task was to identify by forced choice the central vowel of the phrase "La l_la", among 5 possible choices: "a", "é", "i", "u" and "ou" (see Figure 1). The listeners were allowed to replay the stimuli as many times as they wished, although they were encouraged to do so only if absolutely necessary. The test consisted of three consecutive sessions of perceiving (1) AV stimuli, (2) AO stimuli and (3) VO stimuli. This order was the same for all the participants. In contrast, the 120 stimuli within each session were presented in a different random order for each participant.

3. Results

3.1. Vowel identification

Figure 2 shows the average vowel recognition rate in normal and Lombard speech, for the three speech perception modalities. When all vowels are considered, the results show that the average vowel recognition rate was significantly improved in Lombard speech, compared to normal speech (+5.2%, F(1,28)=32.16, $p < .001$), with a similar intelligibility gain for the AV, AO and VO modalities, of +4.2%, +5.7% and +5.6% of recognition respectively (No significant interaction between the factors Speech type and Perceptual modality: F(2,56)=0.309, $p > .7$). Results are more complex within each vowel category (see Figure 3). First, and as expected, a significant interaction was found between the factors Vowel and Perceptual modality (F(8,224)=19.488, $p < .001$). This reflects that some vowels such as [a] and [e] were overall less recognized when auditory information was missing (-14.3% of recognition in the VO modality, compared to the AV and AO modalities), whereas rounded vowels such as [y] and [u] were less recognized in absence of visual information (-30.7% of recognition in the AO modality, compared to AV and VO modalities). Finally, the vowel [i] was very well recognized when both audio and visual information were available, whereas its intelligibility was reduced in the AO and VO modalities (by -35.5% and -27.2%, respectively). A significant interaction between the factors Speech type and Vowel was also found (F(4,112)=11,853, $p < .001$), reflecting the fact that some vowels such as [a] and [i] always tended to be more intelligible in Lombard speech (+15.8% and +12.0% of recognition, respectively) whereas others such as [y] always tended to be more difficult to recognize in Lombard speech (by -5.6%, on average). Listeners showed great variability in vowel identification, such that the intelligibility gain of Lombard speech was statistically significant for the vowel [a] only (F(1,28)=95.863, $p < .001$).

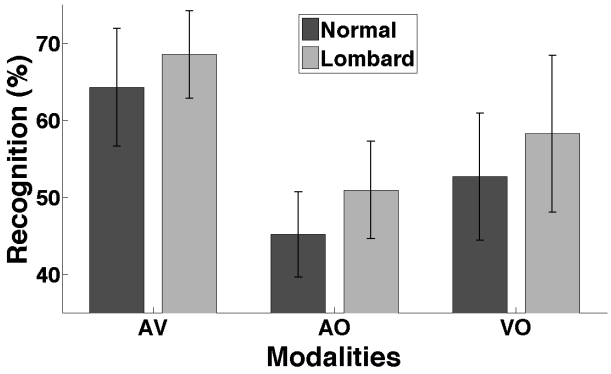

Figure 2: *Average perceptual recognition scores of vowels produced in silence (Normal) and in noise (Lombard), and perceived in the audiovisual (AV), audio-only (AO) and visual-only (VO) modalities.*

3.2. Perceptual recognition and discrimination of phonological features

3.2.1. Vowel height (both audible and visible)

In Figure 4, the first two bars of each graph represent the percentage of cases in which the vowels [i], [y] and [u] were perceived as [i], [y] or [u]. In other words, this percentage represents the recognition score of the phonological feature "closed". The results show that this feature was recognized very robustly in all modalities (mean score of 89.5%). The identification of

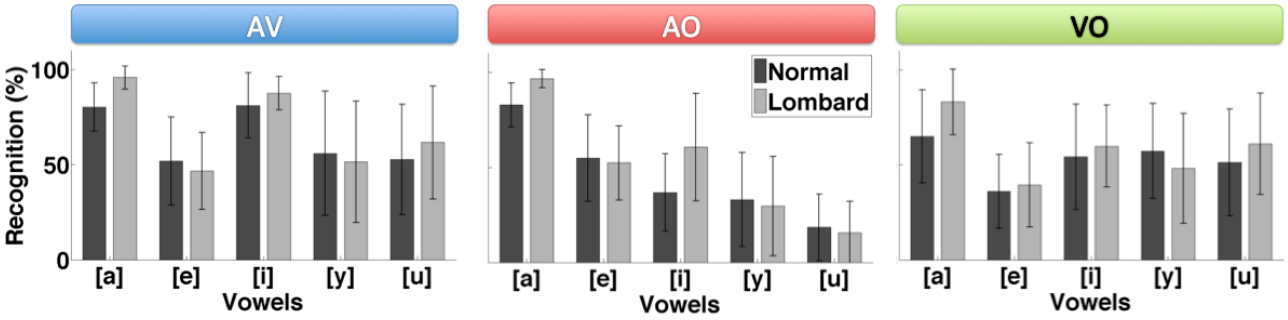

Figure 3: *Perceptual recognition scores of 5 French vowels produced in silence (Normal) and in noise (Lombard), and perceived in the audiovisual (AV), audio-only (AO) and visual-only (VO) modalities. The dotted line indicates the chance level (20%).*

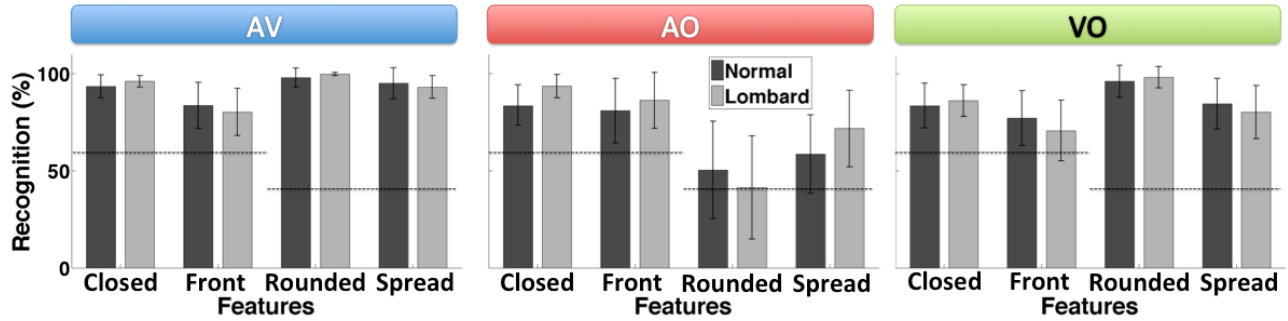

Figure 4: *Scores of correct categorization, by phonological features, of vowels produced in silence (Normal) and in noise (Lombard), and perceived in the audiovisual (AV), audio-only (AO) and visual-only (VO) modalities. The dotted line indicates the chance level (40% for the features "Rounded" and "Spread", 60% for the features "Closed" and "Front").*

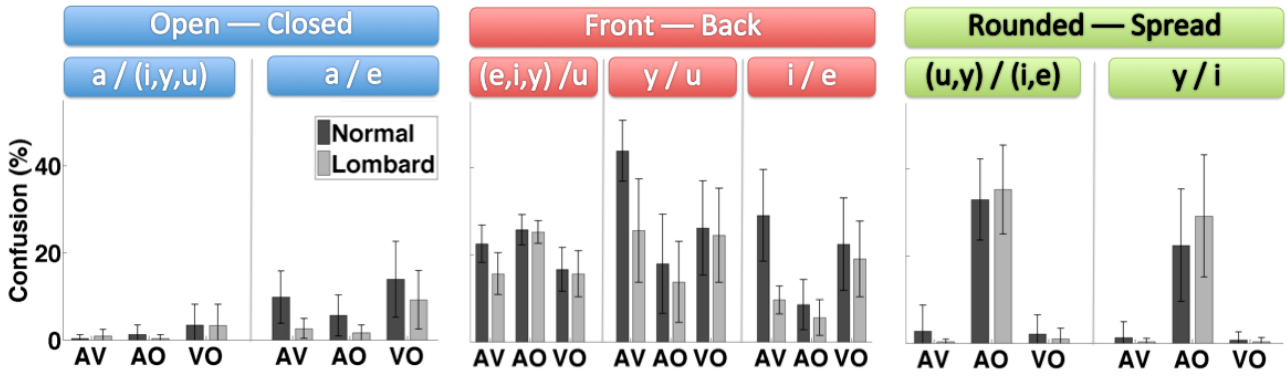

Figure 5: *Confusion scores between vowels or vowel groups that differ in a phonological feature. Vowels were produced in silence (Normal) and in noise (Lombard), and were presented to listeners with the visual modality (VO), the audio modality (AO) or both (AV).*

this feature was significantly improved in Lombard speech for the AO modality (+9.6%), whereas it did not increase much in the two other modalities (+2.6%). (Significant effect of the factor Speech Type: F(1,28)=18,725, $p < .001$ and significant interaction between the factors Speech type and Perceptual modality: F(2,56)=7.725, p=.001). In Figure 5, the first graph represents the percentage of cases in which the opened vowel [a] was perceived as a closed vowel ([i], [y] or [u]), and vice versa. In other words, this percentage represents a misperception score of the "vowel height" feature. Similarly, the second graph represents the confusion score between the open vowel [a] and the mid-closed vowel [e]. The results show that the misperception of the feature "vowel height" was very marginal in all modalities and that the confusion between the vowels [a] and [e] was significantly reduced in Lombard speech (by -5.3%, F(1,28)=46,582, $p < .001$).

3.2.2. Backness (mainly audible)

Figure 4 represents the percentage of cases in which the front vowels [i], [e] and [y] were perceived as a [i], a [e] or a [y]. The middle graphs of Figure 5 represent misperception scores of the feature "backness". The results show that this feature was recognized fairly robustly accross all modalities (average score of 79.8%). A significant interaction was observed between the factors Speech type and Perceptual modality (F(2,56)=14.137,

$p < .001$), showing that the identification of this feature slightly increased in the AO modality (+5.7%) while it decreased in the VO one (-7.8%), finally resulting in an unchanged intelligibility in the AV modality. Furthermore, the results show that confusion between the front vowels [i], [e] and [y] and the back vowel [u] can be relatively high: around 20% in all modalities, with an improvement in Lombard speech for the AV modality. Similarly, confusion between [y] and [u] can be as high as 43.8% in the AV modality. Overall, the misperception of the "backness" feature, which theoretically distinguishes [y] from [u] and [i] from [e], was reduced in the AO modality, compared to the VO modality, and even more reduced compared to the AV modality. Nevertheless, these confusions were significantly reduced in Lombard speech (by -8.3% on average, F(1,28)=109.264, $p < .001$), especially in the AV modality (-18.8%).

3.2.3. Rounding (mainly visible)

Figure 4 represents the percentage of cases in which the rounded vowels [y] and [u] were perceived as a [y] or a [u], and the percentage of cases in which the spread vowels [i] and [e] were perceived as a [i] or [e]. The results show that the "rounded" feature was highly recognized when visual information was available (98.1% of recognition in AV and VO modalities), whereas the recognition rate was not above chance level in the AO modality. Similarly, the "spread" feature was highly recognized in the AV modality (94.2% of recognition) and in the VO modality (82.4%), whereas its identification was significantly more difficult in the AO modality (65.2%, which is above chance level though). However, the two features were not affected in the same way by the Lombard effect. Rounding perception remained unchanged in Lombard speech for VO and AV modalities, whereas it decreased in Lombard speech for the AO modality (-9.1%) (Significant interaction between the factors Speech type and Perceptual modality: F(2,56)=9.687, $p < .001$). On the contrary, Spreading identification tended decrease in Lombard speech for VO and AV modalities (-3.2%), whereas it increased in the AO modality (+13.2%) (Significant interaction between the factors Speech type and Perceptual modality: F(2,56)=12.905, $p < .001$). The last two graphs of Figure 5 represent the confusion rate between the rounded vowels [y] and [u], and the spread vowels [i] and [e], as well as the degree of misperception of the "rounding" feature in the discrimination of the vowels [y] and [i]. The results show that the rounding feature was misperceived only in the AO modality, with a high level of confusion (up to 35%). In contrast to all other confusions, the misperception of the "rounding" feature in the AO modality was significantly increased in Lombard speech (by +4.5% F(1,28)=9.159, p=.005).

4. Discussion and conclusion

At the global word and sentence level, previous studies have shown that Lombard speech is more intelligible than normal speech in both audio and audio-visual modalities [13, 1, 2, 14, 12, 11, 4, 6, 7]. This study confirms this tendency at the more specific level of vowel recognition. Furthermore, previous studies have shown that the intelligibility gain provided by the visual modality, in addition to an auditory-only perception of utterances, was in most cases greater in Lombard speech than in speech produced in silence [15, 11]. This was not the case in our study, focusing specifically on vowel perception, where this gain remained similar in Lombard speech as in conversational speech. In addition, our study specifically examined the consequences of the Lombard effect in the visual modality alone. The results showed that the benefits of the Lombard effect on vowel recognition also extend to the visual modality alone. These different results should be considered with caution, however, since the condition order (AV, AO, VO) was the same for all the participants. We therefore cannot exclude that they might be influenced by a learning or fatigue effect. In any case, such results have clear implications, not only to the improvement of communication efficiency in noisy environnements – especially for people such as teachers, who are confronted daily with such environments in their workplaces [19], but also to the development of speech enhancement techniques [20], and the improvement of the robustness of ASR systems in noisy conditions [2].

The recognition scores observed in this perceptual study can now be discussed at the light of acoustic and articulatory changes observed in production by previous studies of Lombard speech. Vowel height is the only phonological feature that was highly recognized in all modalities, which is logical since the perception of vowel height relies on both audible and visible cues (F1, jaw aperture). Its recognition was improved in Lombard speech, especially in the AO modality. This is consistent with the increased contrast in lip aperture and along the F1-F0 dimension that was previously reported in production for Lombard speech [1, 2, 3, 16, 10]. Rounding was highly recognized in VO and AV modalities, but not beyond the chance level in the AO modality. This also makes sense since the grouping of vowels [y] and [u] in a same phonological category relies primarily on visible cues. Rounding recognition was at its best in normal speech for the VO and AV modalities. It was not reduced in Lombard speech, despite the overall increase in lip opening in noise that has also been observed in rounded vowels by previous production studies [16, 10]. This is consistent with previous studies showing that visual intelligibility of rounded vowels is highly resistant to articulatory changes [21]. On the contrary, audio recognition of rounding decreased in Lombard speech, which is consistent with the significant increase of both F1 and F2 in vowel [u] produced in noise [1, 2, 3, 16, 10]. Similarly, spreading was well recognized in VO and AV modalities. Spread vowels were confused with rounded vowels only in the AO modality. However, the recognition of spreading decreased in Lombard speech in the modalities providing visual information, whereas it increased in the AO modality. The observed degradation in the visual domain is consistent with the reduced contrast in lip spreading that was observed in production for Lombard speech [16, 10]. The improvement observed in the audio domain cannot be explained by variations in F2. It might be found in the variations of F3. Finally, vowel backness was the least well discriminated feature in all perceptual modalities. Most of the confusions between the 4 closed and mid-closed vowels [i], [e], [y] and [u] stemmed from a misperception of the backness feature. This weakness of the backness feature makes sense in the VO modality, since backness relies primarily on audible cues (F2). Its weakness in the AO and AV conditions is more difficult to explain, since these modalities provide the information about F2. Furthermore, perceptual discrimination of front and back vowels was improved in Lombard speech, for all modalities, which is not consistent with the reduced contrast in F2 measured in production for Lombard speech [1, 2, 3, 16, 10].

5. References

[1] 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.

[2] 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.

[3] A. Castellanos, J.-M. Benedí, and F. Casacuberta, "An analysis of general acoustic-phonetic features for spanish speech produced with the lombard effect," *Speech Communication*, vol. 20, no. 1-2, pp. 23–35, 1996.

[4] Y. Lu and M. Cooke, "The contribution of changes in f0 and spectral tilt to increased intelligibility of speech produced in noise," *Speech Communication*, vol. 51, no. 12, pp. 1253–1262, 2009.

[5] M. Garnier, N. Henrich, and D. Dubois, "Influence of sound immersion and communicative interaction on the lombard effect," 2010.

[6] J. Kim, A. Sironic, and C. Davis, "Hearing speech in noise: Seeing a loud talker is better," *Perception*, vol. 40, no. 7, pp. 853–862, 2011.

[7] S. Alexanderson and J. Beskow, "Animated lombard speech: motion capture, facial animation and visual intelligibility of speech produced in adverse conditions," *Computer Speech & Language*, vol. 28, no. 2, pp. 607–618, 2014.

[8] M. Fitzpatrick, J. Kim, and C. Davis, "The effect of seeing the interlocutor on auditory and visual speech production in noise," *Speech Communication*, vol. 74, pp. 37–51, 2015.

[9] J. Šimko, Š. Beňuš, and M. Vainio, "Hyperarticulation in lombard speech: Global coordination of the jaw, lips and the tongue," *The Journal of the Acoustical Society of America*, vol. 139, no. 1, pp. 151–162, 2016.

[10] M. Garnier, L. Ménard, and B. Alexandre, "Hyper-articulation in lombard speech: An active communicative strategy to enhance visible speech cues?" *The Journal of the Acoustical Society of America*, vol. 144, no. 2, pp. 1059–1074, 2018.

[11] E. Vatikiotis-Bateson, A. V. Barbosa, C. Y. Chow, M. Oberg, J. Tan, and H. C. Yehia, "Audiovisual lombard speech: reconciling production and perception." in *AVSP*, 2007, p. 41.

[12] V. Chung, N. Mirante, J. Otten, and E. Vatikiotis-Bateson, "Audiovisual processing of lombard speech." in *AVSP*. Citeseer, 2005, pp. 55–56.

[13] J. J. Dreher and J. O'Neill, "Effects of ambient noise on speaker intelligibility for words and phrases," *The Journal of the Acoustical Society of America*, vol. 29, no. 12, pp. 1320–1323, 1957.

[14] A. L. Pittman and T. L. Wiley, "Recognition of speech produced in noise," 2001.

[15] C. Davis, J. Kim, K. Grauwinkel, and H. Mixdorff, "Lombard speech: Auditory (a), visual (v) and av effects," in *Proceedings of the third international conference on speech prosody*. Citeseer, 2006, pp. 248–252.

[16] M. Garnier, "May speech modifications in noise contribute to enhance audio-visible cues to segment perception?" in *AVSP*. Citeseer, 2008, pp. 95–100.

[17] G. Turner, L. Roach, and R. de Jonge, "Lip contact pressure while talking in background noise," *Perspectives of the ASHA Special Interest Groups*, vol. 1, no. 19, pp. 5–14, 2016.

[18] W. H. Sumby and I. Pollack, "Visual contribution to speech intelligibility in noise," *The journal of the acoustical society of america*, vol. 26, no. 2, pp. 212–215, 1954.

[19] L. M. Rantala, S. Hakala, S. Holmqvist, and E. Sala, "Classroom noise and teachers' voice production," *Journal of Speech, Language, and Hearing Research*, vol. 58, no. 5, pp. 1397–1406, 2015.

[20] M. D. Skowronski and J. G. Harris, "Applied principles of clear and lombard speech for automated intelligibility enhancement in noisy environments," *Speech Communication*, vol. 48, no. 5, pp. 549–558, 2006.

[21] C. Benoit, T. Mohamadi, and S. Kandel, "Effects of phonetic context on audio-visual intelligibility of french," *Journal of Speech, Language, and Hearing Research*, vol. 37, no. 5, pp. 1195–1203, 1994.