Analysis of Lombard speech prosody using WaveNet embeddings

Juraj Šimko & Antti Suni
University of Helsinki

When we speak in a loud environment, we systematically increase the intensity and pitch, and change voice quality characteristics. This transformation, called Lombard effect, is an adaptation of speech to noise in order to increase the signal-to-noise ratio [1]. The precise realization of the adaptation is known to depend on the noise level, noise type and linguistic content [2, 3].

We present a novel methodology for analysis of quantitative and qualitative nature of Lombard speech and its interaction with focus structure of utterances. The method is based on the auto-regressive, convolutional WaveNet speech synthesis system [4] trained on a corpus of Lombard speech with Finnish utterances varying in noise type and its level as well as in focus structure (3 focus types; see [5] for the full description). The corpus recordings were downsampled and low-pass filtered with only prosodic characteristics of the original signal (such as f_0 and energy) thus preserved.

The influence of the noise and focus type was captured in the form of a *conditioning embedding* that was trained as a part of the WaveNet synthesizer. As the training material contained utterances from 21 speakers, another embedding was used to disentangle the effect of speaker on the synthesis. We will show that the resulting representation of the noise type, noise level and focus structure captures the effects of all of these influences on speech – as well as interactions among them – in an intuitively interpretable way. We will also compare the results of our analysis with the results of more traditional examination of the same corpus presented in [5].

References

[1] E. Lombard, “Le signe de l’elevation de la voix,” *Ann. Maladies Oreille, Larynx, Nez, Pharynx*, vol. 37, no. 101-119, p. 25, 1911.

[2] R. Patel and K. Schell, “The influence of linguistic content on the Lombard effect,” *Journal of Speech, Language, and Hearing Research*, vol. 51, no. 1, p. 209, 2008.

[3] C. Riversand, M. Rastatter, “The effects of multitalker and masker noise on fundamental frequency variability during spontaneous speech for children and adults.” *Journal of Auditory Research*, 1985.

[4] A. v.d. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu. “Wavenet: A generative model for raw audio.” *arXiv preprint arXiv:1609.03499*, 2016.

[5] M. Vainio, D. Aalto, A. Suni, A. Arnhold, T. Raitio, H. Seijo, J. Järvikivi, and P. Alku. “Effect of noise type and level on focus related fundamental frequency changes.” In *Thirteenth Annual Conference of the International Speech Communication Association*, 2012.