Interspeech 2025

Rotterdam, 17 - 21 August 2025

Place:

Rotterdam, Netherlands

Title:

Gradual modeling of the Lombard effect by modifying speaker embeddings from a Text-To-Speech model

Authors:

Thiago Lobato; Magnus Schäfer

Abstract:

The Lombard effect is defined as the involuntary adaptation speakers go through when strong background noise is present. Depending on the speaker, the properties of their speech, such as level, pitch and duration, may drastically change. Thus, having Lombard-compatible samples is essential for a full evaluation of communication systems in which speech is in the foreground and strong background noise may be present, such as in-car communication systems or phone-calls in a loud environment. It is possible to use real recordings of Lombard speech for some cases, but this would increase the necessary recording time – particularly if different background noise scenarios should be considered. Additionally, there are cases where existing signals are obligatory for certain tests and recording additional material is not feasible at all.

This work proposes to generate speech in a “Lombard version” of the original voice as an alternative. The Lombard effect is modeled by modifying speaker embeddings used as conditioning to a text-to-speech model, an auto-regressive GPT transformer for Encodec tokens with a multi-band diffusion decoder. The transformation to Lombard speech is done using a feedforward neural network with SwiGLU activations which is learned based on embeddings of plain and Lombard speech pairs presented in the Audio-Visual Lombard Grid Speech corpus. Finally, the signal level is increased according to the methodology used in ITU-T P.1150 and a neural vocoder is used for voice stretching. This approach generates speech that possesses all relevant properties of the Lombard speech while retaining the identity of the original speaker. Additionally, by optimally interpolating the embeddings, we can generate gradual levels of Lombard speech. We then show that the model can be used to generate Lombard variations of existing signals (e.g. the speech sequences from ITU-T P.501) for different background noise levels. The quality of the results is evaluated with a MOS score for speaker similarity and speaker naturalness.

HEAD acoustics GmbH
Ebertstraße 30a
52134 Herzogenrath, Germany

Thiago Lobato
Mail: info@head-acoustics.com
Telephone: +49 2407 577-0