Examining the Self-Similarity Method for the Lombard Effect Recognition

Gražina Korvel[1], Krzysztof Kąkol[2], Povilas Treigys[1], and Bożena Kostek[3]

[1]Institute of Data Science and Digital Technologies, Vilnius University, Vilnius, Lithuania
[2]GPS Software, Gdansk, Poland
[3]Audio Acoustics Laboratory, Faculty of Electronics, Telecommunications and Informatics, Gdansk University of Technology, Gdansk, Poland

Introduction

Well-known phenomena in the signal included in the Lombard speech are the following: the increased volume of the uttered speech, fundamental frequency rise, formant frequency rise, spectral tilt, duration of utterances, prosody alteration. Most of these features can easily be determined, but observing changes in these features in the context of the Lombard speech is not so simple. The main reason for this is that the Lombard speech characteristics vary according to the noise level. **In this research, the self-similarity method is employed for the Lombard effect recognition in the presence of noise. Self-similarity matrices based on acoustic parameters related to the Lombard effect are created and introduced as 2D space features at the CNN input.**

An example of a similarity matrix

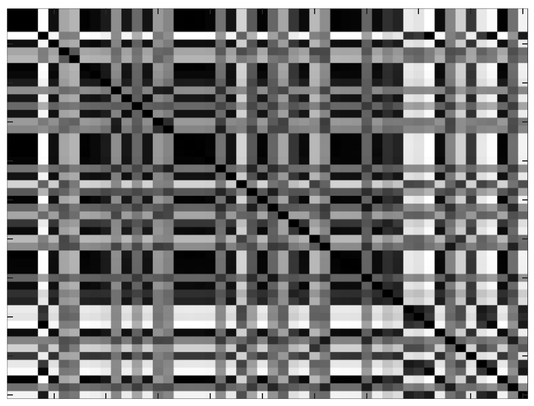

Fig. 1. The similarity matrices of the utterance **without the Lombard effect**

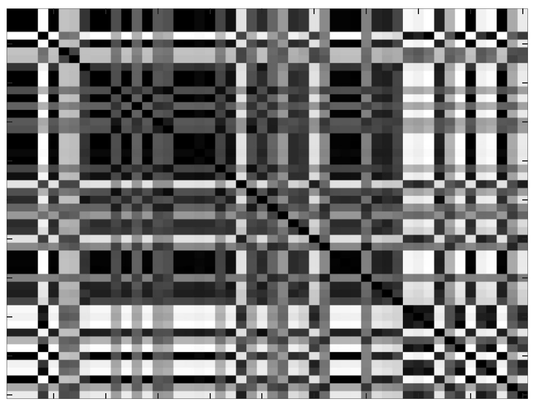

Fig. 2. The similarity matrices of the same utterance **with the Lombard effect**

Similarity matrix construction

Let p_i and p_j be two vectors of parameters:

$$\boldsymbol{p}_i = (p_{i1}, p_{i2}, \ldots, p_{iN})$$
$$\boldsymbol{p}_j = (p_{j1}, p_{j2}, \ldots, p_{jN}),\ i, j \in [1, M]$$

Where N denotes the number of short-time intervals, M - the number of parameters

The similarity matrix was constructed from the pairwise distances between parameters, calculated by the following formula:

$$d(\boldsymbol{p}_i, \boldsymbol{p}_j) = \sqrt{\sum_{n=1}^{N} (p_{in} - p_{jn})^2}$$

The acoustic parameters employed:

- Peak to RMS
- Audio Spectral Kurtosis
- Audio Spectrum Envelope calculated on 29 sub-bands
- Mean Spectral Flatness Measure
- Mel-Frequency Cepstral Coefficients

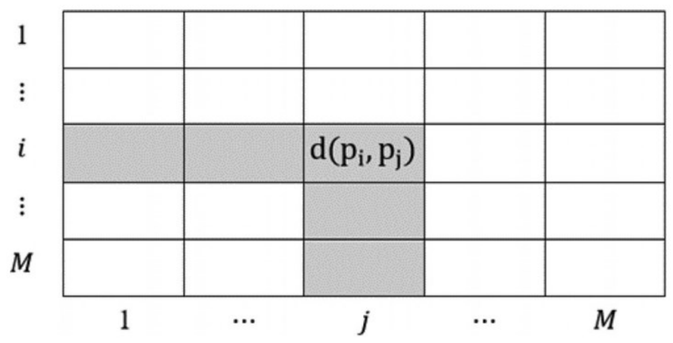

Fig. 3. A graphical representation of the similarity matrix construction

Experimental setup

The experiments were performed on recordings of 8 speakers. The normal speech utterances were recorded without additional noise played back. The utterances with the Lombard effect were achieved by playing interference pink noise via the headphones during the recording process. The recording scenario included 15 sentences and was repeated twice (in two rooms with different acoustic characteristics).

Step 1. Dividing a speech signal into short-term segments (1024 samp.)
Step 2. Extraction of acoustic parameters
Step 3. Dividing a speech signal into mid-term segments (40 short-term segments)
Step 4. Construction of a similarity matrixes

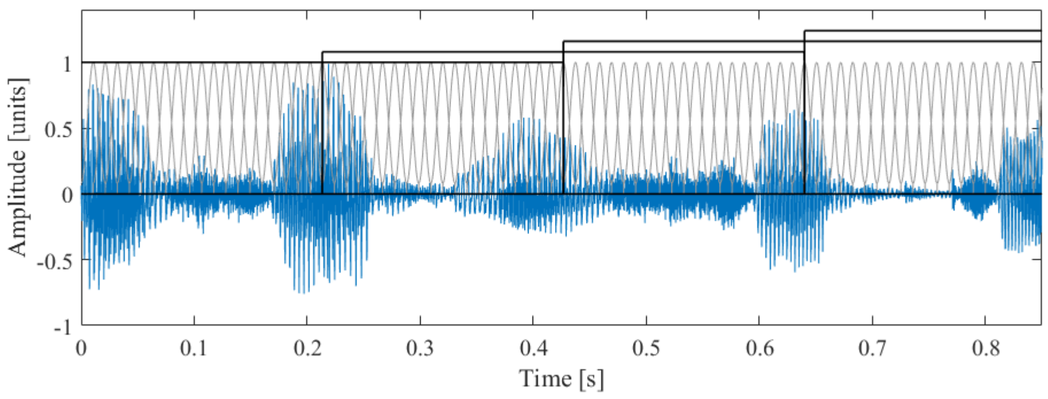

Fig. 4. An example of the dividing a speech signal into short-term and mid-term segments

Experimental Results

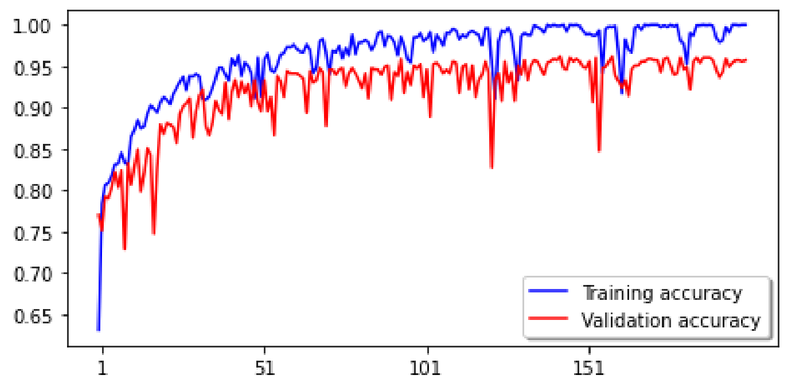

Fig. 5. Training and validation accuracy (CNN Model: number of filters = 8, filter size = 3, pool size = 2)

Test accuracy: 93.52 %

- ✓ *Total train images: 1468, validate images: 630, test images: 525.*
- ✓ *Epochs: 200.*

Conclusions

The self-similarity-based method showed promising results in highlighting acoustic differences between normal and Lombard speech.

Acknowledgment

This research is funded by the European Social Fund under the No 09.3.3-LMT-K-712 “Development of Competences of Scientists, other Researchers and Students through Practical Research Activities” measure.