LITHUANIAN COMPUTER SOCIETY
VILNIUS UNIVERSITY
INSTITUTE OF DATA SCIENCE AND DIGITAL TECHNOLOGIES
LITHUANIAN ACADEMY OF SCIENCES

11th International Workshop on

DATA ANALYSIS METHODS FOR SOFTWARE SYSTEMS

Druskininkai, Lithuania, Hotel "Europa Royale"
http://www.mii.lt/DAMSS

November 28–30, 2019

VILNIUS UNIVERSITY PRESS
Vilnius, 2019

Co-Chairmen:

Dr. Saulius Maskeliūnas (Lithuanian Computer Society)
Prof. Gintautas Dzemyda (Vilnius University, Lithuanian Academy of Sciences)

Programme Committee:

Prof. Juris Borzovs (Latvia)
Prof. Albertas Čaplinskas (Lithuania)
Prof. Robertas Damaševičius (Lithuania)
Prof. Janis Grundspenkis (Latvia)
Prof. Janusz Kacprzyk (Poland)
Prof. Ignacy Kaliszewski (Poland)
Prof. Yuriy Kharin (Belarus)
Prof. Tomas Krilavičius (Lithuania)
Prof. Julius Žilinskas (Lithuania)

Organizing Committee:

Dr. Jolita Bernatavičienė
Prof. Olga Kurasova
Dr. Viktor Medvedev
Laima Paliulionienė
Dr. Martynas Sabaliauskas

Contacts:

Dr. Jolita Bernatavičienė
jolita.bernataviciene@mif.vu.lt
Prof. Olga Kurasova
olga.kurasova@mif.vu.lt
Tel. +370 5 2109 315

https://doi.org/10.15388/Proceedings.2019.8

ISBN 978-609-07-0325-0 (digital PDF)

Distortion-Based Audio Augmentation for Continuous Speech Recognition

Jūratė Vaičiulytė, Gintautas Tamulevičius

Institute of Data Science and Digital Technologies
Vilnius University
jurate.vaiciulyte@mif.vu.lt

The amount of training data is an important issue in continuous speech recognition, both neural networks and statistical approaches. Insufficient or inconsistent training data set (which is the case for low-resource languages) may give low quality or over-trained models resulting in poor speech recognition. Nowadays, the principle of artificially increasing the amount of training is widely applied. The amount of data is increased by adding extra noise, modifying time scale, perturbating the speaker's vocal tract length, distorting acoustics features, or applying speech synthesis to obtain additional data.

In this study, we have explored the influence of distortion-based speech data augmentation on continuous speech recognition rate and its robustness. For this purpose, we have employed additive and convolutional noises for the speech recordings (the speech corpus of ~90 hrs was exploited). The white noise was added at various levels to form an additive noise-based subset of the training data. The convolutional noise was imitated with the help of various room impulse responses (we have used the BUT Speech@FIT Reverb Database for this purpose). The amount of distorted data in the training set was formed by randomly selecting utterances and consistently increasing their amount, thus analysing their impact on recognition accuracy and robustness.

The effect of augmented training data on speech recognition was compared to a corpus of ~200 hrs, advantages and disadvantages of each case were analysed.