Acoustic analysis of pathologic voice: What is done and what is next?

G. Tamulevičius[1], N. Šiupšinskienė[2], M. Danilovaitė[1]

[1] Institute of Data Science and Digital Technologies, Vilnius University
[2] Department of Otolaryngology, Hospital of Lithuanian University of Health Sciences Kaunas Clinics

INTRODUCTION

Acoustic analysis-based pathologic voice assessment is not a new task and research activity. Sixty years ago, researchers focused on the changes in the voice generation process, the physiological causes of these changes, and the acoustic features of pathologies. Nowadays, artificial intelligence-based approaches dominate studies: end-to-end pathologic voice analysis, intelligent feature selection techniques, and evaluation of pathology degrees.

In this study, we present the results of the acoustic analysis of the pathologic voice review. We have analyzed studies starting from the first results until recent ones, summarized and grouped these results by decades. We have tried highlighting the critical achievements and trending directions in pathological voice analysis: features and feature sets, feature selection, and decision-making.

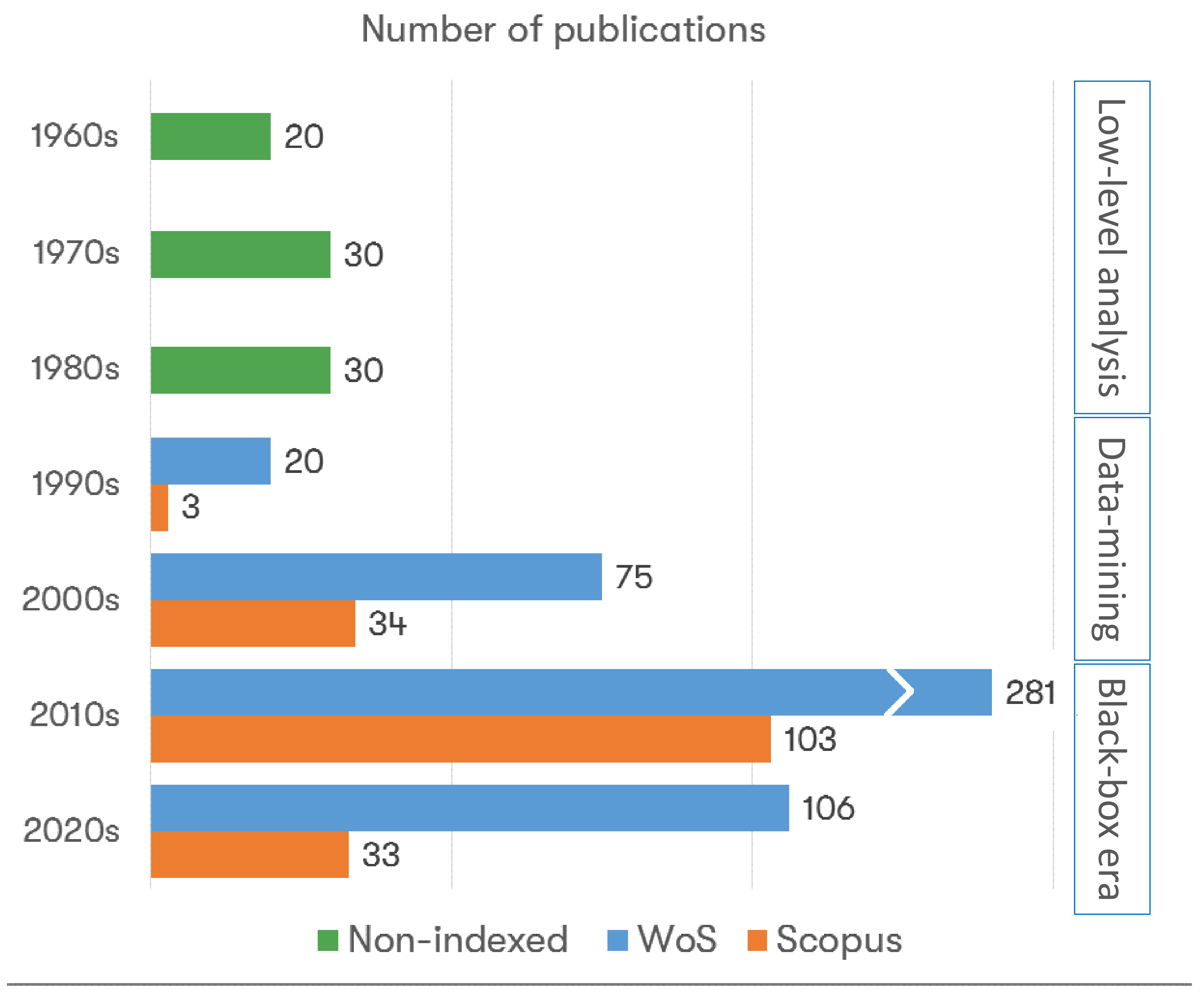
WHAT IS DONE

The main achievements and ideas in the light of decades:

- **1960s**: Pathologic voice analysis was based on knowledge about normal and abnormal fold functioning, therefore entire acoustical analysis was based on time-domain techniques. The goal of the acoustic analysis was to detect and evaluate perturbation of pitch and amplitude which were considered attributes of pathology. Dominant features: time difference between adjacent pitch periods, perturbation factors of pitch and amplitude.
- **1970s**: The autoregressive model (LPC also) comes to the scene, it enables inverse filtering of the speech signal and glottis signal extraction. Low-level acoustic features (with new names of jitter and shimmer) still dominate in pathologic voice analysis, and spectral shape features (like spectral flatness, spectral envelope) show up.
- **1980s**: The concept of cepstra is introduced alongside the concept of spectra: cepstral features are proposed. Widely used time-domain features are given their own names (PPQ, APQ, jitter, shimmer). The idea of evaluating the noise in the speech signal is introduced (NNE, HNR).
- **1990s**: The CPP (Cepstral Peak Prominence) feature is proposed. The best-known *Saarbruecken Voice Database* is created (still used actively). Researchers still use time domain features, but cepstral and spectral features are also used quite intensively.
- **2000s**: New features are proposed actively (e.g., MFCC, LPC cepstral features). Heterogeneous feature sets are created (by combining different features), complemented by functionals of low-level features. As the size of the feature sets grows, various feature selection methods are introduced. The first applications of the artificial neural network are presented.
- **2010s**: Data mining and machine learning techniques dominate pathological voice analysis: large feature sets are implemented, and feature dimensionality reduction is applied. The *OpenSmile* tool is also used for multiple feature generation. The relationship between subjective and objective voice assessment is analyzed.
- **2020s**: The "black box" paradigm dominates: huge joint feature sets are analyzed using deep learning techniques. Their results are evaluated by correlation with the human expert's diagnoses. The main research questions relate to the architecture and topology of the deep learning network. The same already-known features dominate the analysis. Most of the proposed approaches still try to solve the pathologic voice detection problem.

Still, the question of pathologic voice detection, pathology identification is open.

INTENSITY OF RESEARCH



- The publications are indexed in databases starting from the 1990s. Until the 1990s, the number of publications corresponded to those discovered publicly.
- The publication search condition: ("acoustic analysis pathologic voice") OR ("acoustic analysis voice pathology") OR ("acoustic analysis fold pathology") OR ("acoustic analysis pathologic folds") OR ("acoustic analysis vocal folds").
- The discovered publications were limited to acoustics, engineering, multidisciplinary, computer science, applications domains.

SKEPTICISM

- No essentially new ideas in the last few decades, some signs of research stagnation or cyclicality can be seen. E.g., acoustic features proposed and studied 60 years ago make a comeback today with a deep networks-based classification.
- The applied features are not grounded, and the relationship with the physical attributes of pathologies is unclear.
- The "black box" and classification paradigms make the problem of type "TRUE/FALSE", it does not give any physically grounded result.
- There is no objective relationship between classification results and subjective assessment techniques practiced by clinicians.
- Most of the explored datasets are private and inaccessible for public use. Therefore, the comparison of studies, experience, and knowledge sharing is limited.

WHAT IS NEXT?

- One day the "saturation" level of the deep learning techniques will be reached…
- Clinicians still need objective tools related to their subjective techniques. They do not need an automated diagnosis system.
- The nature and characteristics of the acoustic speech signal should be remembered. The speech signal contains most of the information needed for diagnosis.
- Some new ideas? Required.

Vilnius University
Faculty of Mathematics and Informatics
Institute of Data Science and Digital Technologies