# Synthesis and Classification of Amplitude Modulated Glottal Area Waveforms Observed in Vocal Fry

Vinod Devaraj, Philipp Aichinger

Department of Otorhinolaryngology, Division of Phoniatrics-Logopedics,
Medical University of Vienna, Austria

## Introduction

- Characterization of voice quality plays an important role in clinical care for the diagnosis of disordered voices.
- Voice quality types: rough, breathy, vocal fry, diplophonic, hoarse, …
- Vocal fry - voice quality characterized by distinctly audible cycles
  - (one of the types of creaky voice [1])
- Amplitude modulated vocal fry glottal area waveforms (GAWs) without long closed phase [2].

**Objective:**

- To propose a synthesis model for amplitude modulated vocal fry GAWs.
- To distinguish the amplitude modulated vocal fry GAWs from other GAWs of other voice types by detecting modulations.

## Data collection

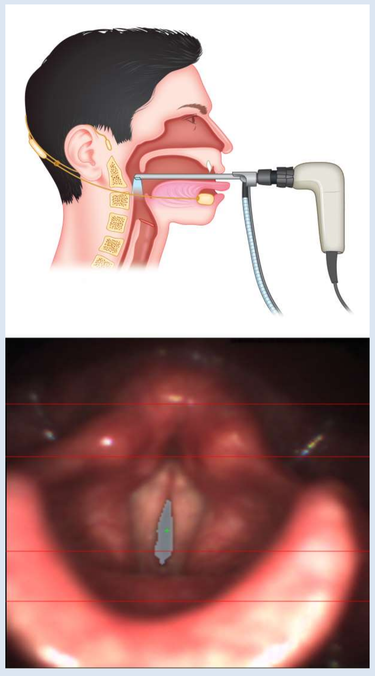


Figure 1: GAW extraction

- Recorded laryngeal high-speed videos (HSVs).
- Extracted GAWs by image segmentation of the HSVs using DNNs.
- GAW - a time series of the projected area of the space between vocal folds during phonation.
- Data used in this study:

| GAWs | Normal | Vocal Fry | | | Diplophonic |
|---|---|---|---|---|---|
| | | CAMS-2 | CAMS-4 | AAMS | |
| Natural | 8 | 2 | 2 | 3 | 29 |
| Synthetic | 300 | 300 | 300 | 300 | 60 |

- Positive group: Vocal fry GAWs
- Negative group: Normal (“euphonic”) and diplophonic GAWs

## Methods

**Block Diagram :**

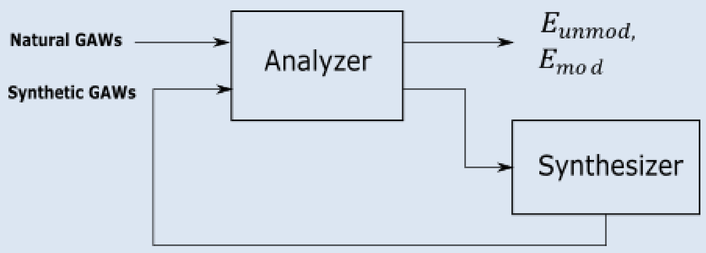



Figure 2: Overview of the analyzer and the synthesizer

Analyzer:

- Analyzer models: 1) modulating 2) non-modulating
- Model GAWs are fitted to the input GAWs by the two models using an analysis-by-synthesis approach [3].
- Pulse trains synthesized by the two models reflect the locations of the extrema of the extrema of the input GAW.
- Jitter and shimmer: fast frequency and amplitude modulations
- A Fourier synthesizer models the input GAWs.
- Model GAWs are optimized by improving the root mean square error $E$ via adding pulse-to-pulse modulations.

Modelling error:

$$E = 20 \cdot log_{10}\left(\frac{\sqrt{\overline{(g-\hat{g})^2}}}{\sqrt{\overline{g^2}}}\right) \; [dB]\,,$$

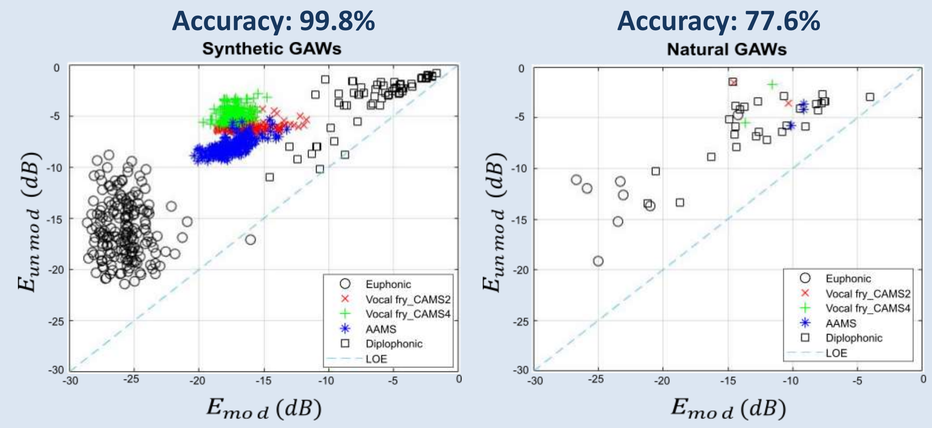
where $g$ is the input GAW and $\hat{g}$ is the model GAW.

Synthesizer:

- Parameters of pulse shape [4] and modulations of the input GAWs are extracted by the analyzer.
- GAWs are synthesized using the distributions of the parameters obtained by the analyzer [5].
- $E_{mod}$ and $E_{unmod}$ are output by the analyzer (modulating and unmodulating model).
- $E_{mod}$ and $E_{unmod}$ are used as predictors by a quadratic SVM classifier.

## Results

- The use of the modulating model in analysis-by-synthesis results in smaller modelling errors.
- Scatter plots of $E_{mod}$ and $E_{unmod}$ enable distinction of voice quality types.
- Natural and synthetic normal GAWs are well separated from the vocal fry GAWs types in the feature space.
- Modelling errors of the natural diplophonic GAWs overlap with the modelling errors of the vocal fry GAWs.
- Distinction is better for synthetic than for natural




(a) (b)

Figure 3: Modelling errors of the GAWs (a) Synthetic and (b) Natural

## Discussion

- Proposed a modulating model for amplitude modulated vocal fry GAWs without closed phase.
- Classification accuracy of natural vocal fry GAWs from diplophonic should be improved.
- More features could be used to distinguish the vocal fry and diplophonic GAWs.
- Residual mismatch between natural and synthetic GAWs’ properties should be reduced → Classical example of bias.

## References and acknowledgements

[1] P. Keating *et al,* “Acoustic properties of different kinds of creaky voice.” ICPhS, 2015.
[2] V. Devaraj *et al*, “A glottal area waveform model for multi-pulsed vocal fry.” MAVEBA, 2019.
[3] P. Aichinger *et al*, “Detection of extra pulses in synthesized glottal area waveforms of dysphonic voices.” *Biomedical signal processing and control* 50 ,158-167, 2019.
[4] G. Chen *et al*, “Estimating the voice source in noise.” Thirteenth Annual Conference of the International Speech Communication Association, 2012.
[5] H. Purwins *et al*, “Deep learning for audio signal processing.” *IEEE Journal of Selected Topics in Signal Processing* 13.2: 206-219, 2019.

This work was supported by the Austrian Science Fund (FWF), KLI722-B30 and the University Hospital Erlangen kindly provided the segmentation tool.