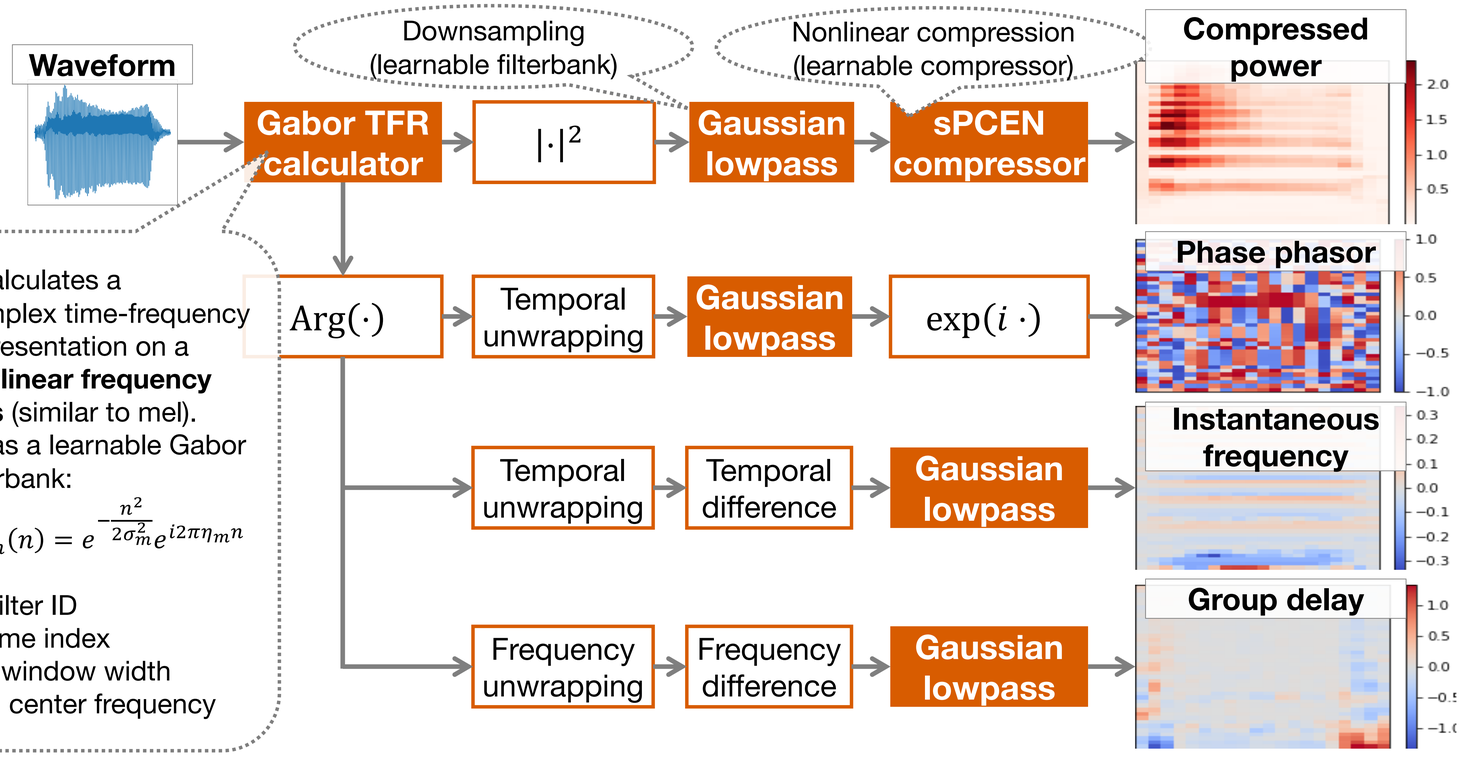
ICASSP 2022
MLSP-21.5

AN INVESTIGATION OF THE EFFECTIVENESS OF PHASE FOR AUDIO CLASSIFICATION

Shunsuke Hidaka[1], Kohei Wakamiya[2], Tokihiko Kaburagi[2]

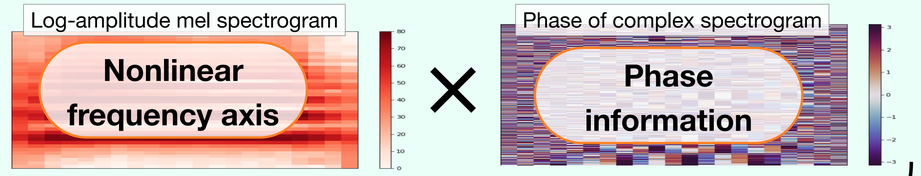
[1] Graduate School of Design, Kyushu University, Fukuoka, Japan [2] Faculty of Design, Kyushu University, Fukuoka, Japan

LEAF-extended (LEarnable Audio Frontend - extended)



Background

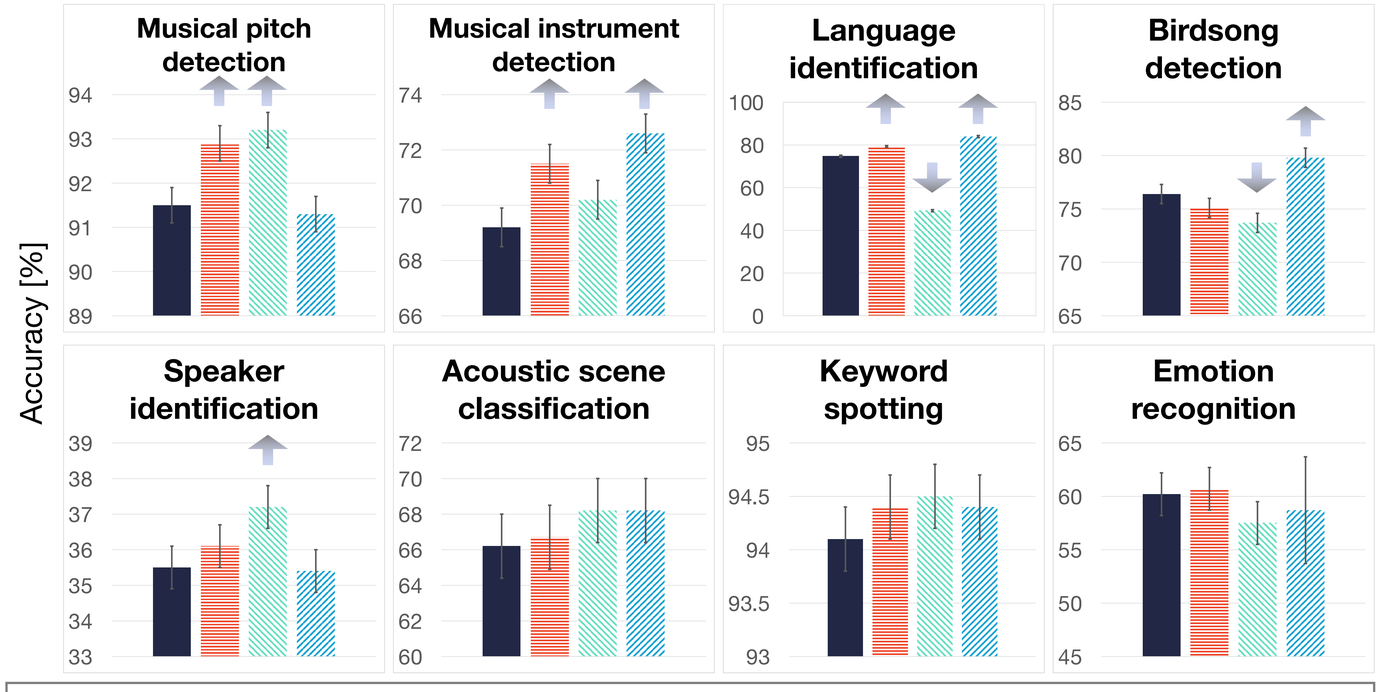
- For speech recognition and audio classification tasks, the **log-amplitude mel spectrogram** has been widely used.
- For extracting clean waveforms from a mixed waveform, such as source separation, features including **phase information** (raw waveforms and complex spectrograms) are currently the mainstream.



Is the combination of both characteristics more effective for audio classification?

Methods

- We propose a learnable audio frontend, LEAF-extended.
- LEAF-extended can calculate **power and phase features** on a **nonlinear frequency axis** (similar to the mel filterbank before learning).
  - **Compressed power**
  - **Phase phasor**
  - **Instantaneous frequency**, which is the time derivative of the phase
  - **Group delay**, which is the frequency derivative of the phase



Experimental Results

- We compared the performance of features in **eight audio classification tasks** using a CNN (EfficientNetB0, about 4 million parameters) connected to LEAF-extended.
- Compared to using the compressed power alone,
  - the performance **improved in three tasks** by adding the **phase phasor**.
  - the performance **improved in two tasks** by adding the **instantaneous frequency**.
  - the performance **improved in three tasks** by adding the **group delay**.
  - the performance *degraded in two tasks* by adding the *instantaneous frequency*.

Discussion and Conclusion

- The addition of the phase information did not necessarily improve performance.
  The phase information could conversely interfere with learning, depending on the dataset.
  In the tasks where the instantaneous frequency caused performance degradation, it was presumed that the recording environments were overfitted.
- For a specific task, if the phase phasor significantly improved performance, then the derivatives of the phase (the instantaneous frequency or group delay) always significantly improved performance as well.
  This fact suggests that in audio classification, the relationship between adjacent elements of the phase is more important than the phase value itself.