AN INVESTIGATION OF THE EFFECTIVENESS OF PHASE FOR AUDIO CLASSIFICATION

ICASSP 2022

MLSP-21.5

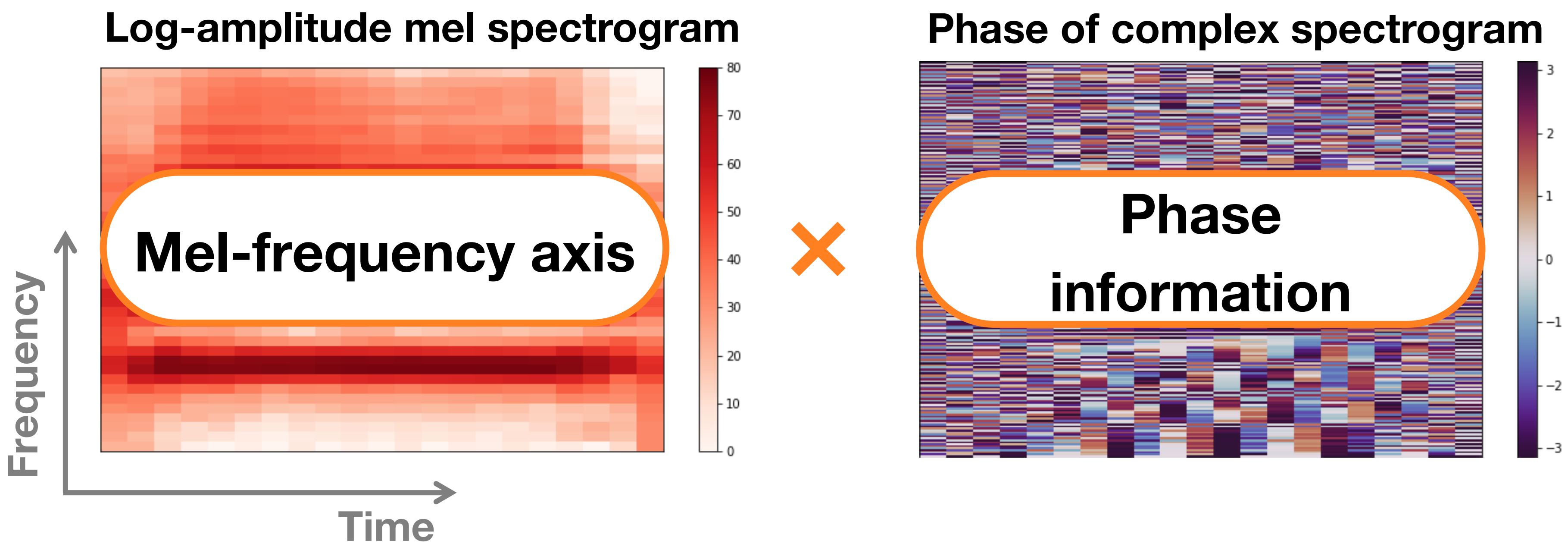
Shunsuke Hidaka[1], Kohei Wakamiya[2], Tokihiko Kaburagi[2]

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

[2] Faculty of Design, Kyushu University, Fukuoka, Japan

1 Minute Summary

- The **log-amplitude mel spectrogram** has widely been used in many tasks.
- The effectiveness of **phase information** was shown recently in tasks such as speech enhancement and source separation.



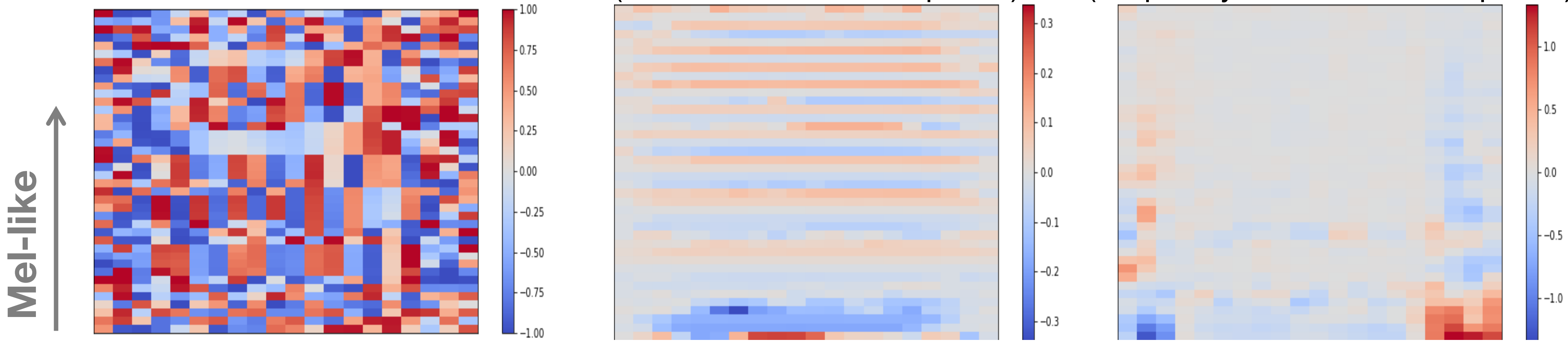
1 Minute Summary

- The **log-amplitude mel spectrogram** has widely been used in many tasks.
- The effectiveness of **phase information** was shown recently in tasks such as speech enhancement and source separation.
- We propose a learnable audio frontend that can calculate the **phase and its derivatives on a mel-like frequency axis**.
- This study investigated the effectiveness of the phase features in eight audio classification tasks.

Phase phasor

Instantaneous frequency
(time derivative of the phase)

Group delay
(frequency derivative of the phase)



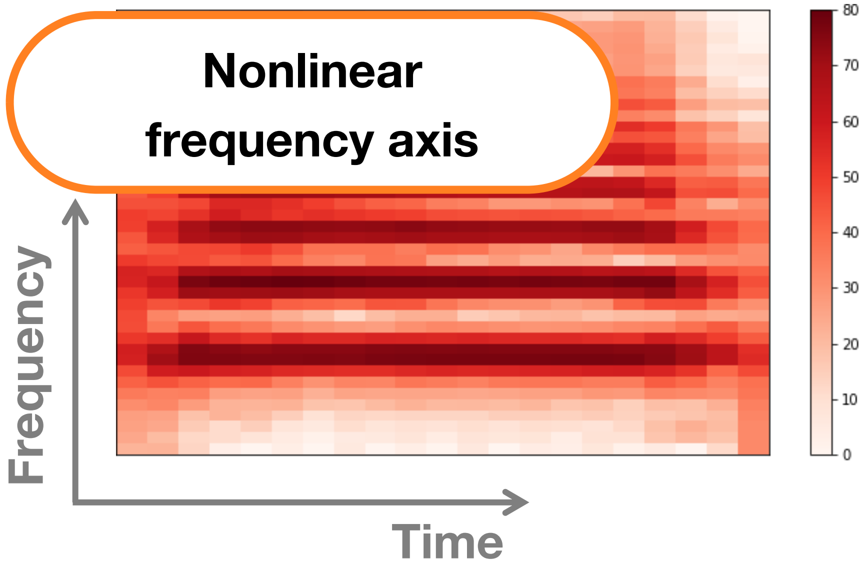
1 Minute Summary

- The **log-amplitude mel spectrogram** has widely been used in many tasks.
- The effectiveness of **phase information** was shown recently in tasks such as speech enhancement and source separation.
- We propose a learnable audio frontend that can calculate the **phase and its derivatives on a mel-like frequency axis**.
- This study investigated the effectiveness of the phase features in eight audio classification tasks.
- The experimental results showed that the phase features significantly **improved performance in five tasks**.
- In contrast, **overfitting to the recording environments** was observed in two tasks.
- The results implied that the relationship between the phase values of adjacent elements is more important than the phase itself in audio classification.

Introduction: Mel-Frequency Feature Representation

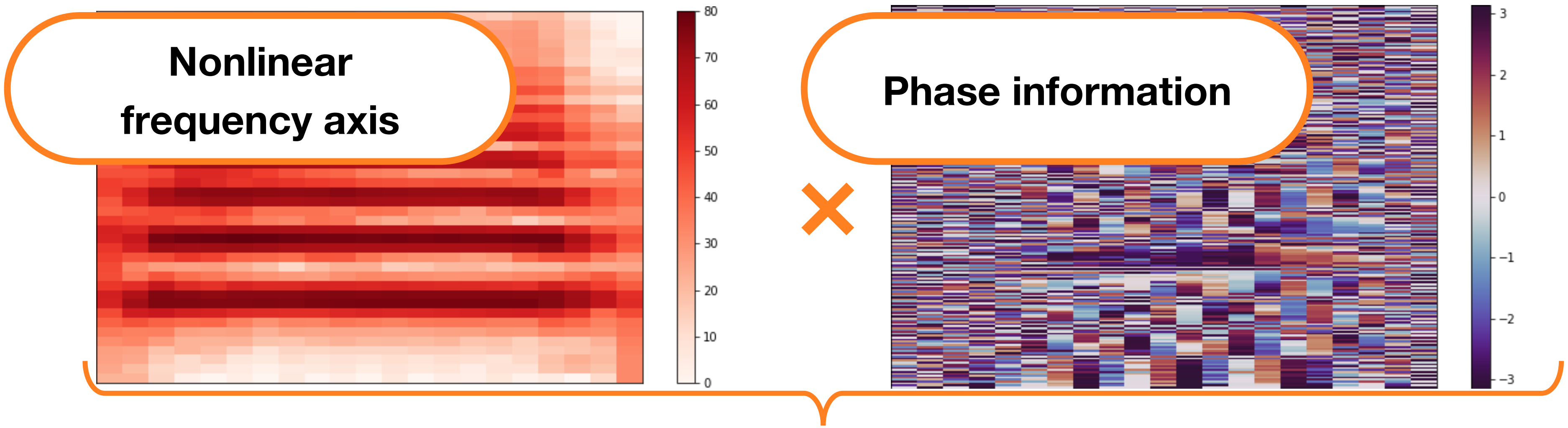
- **Log-amplitude mel spectrogram**
  - is used for audio classification, speech recognition, etc. [Zhang+2020, Heittola+2020]
- **Features including phase information**
  - are such as complex spectrograms and raw waveforms.
  - are used for speech enhancement, source separation, etc. [Luo+2019, Hu+2020]

Log-amplitude mel spectrogram



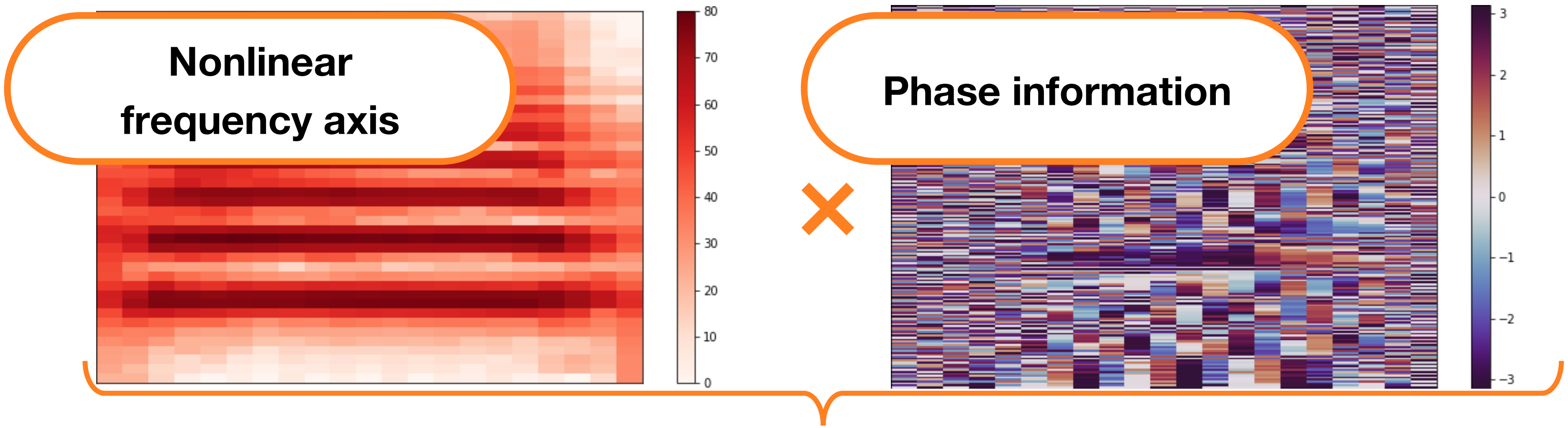
Phase of complex spectrogram

Introduction: Mel-Frequency Feature Representation

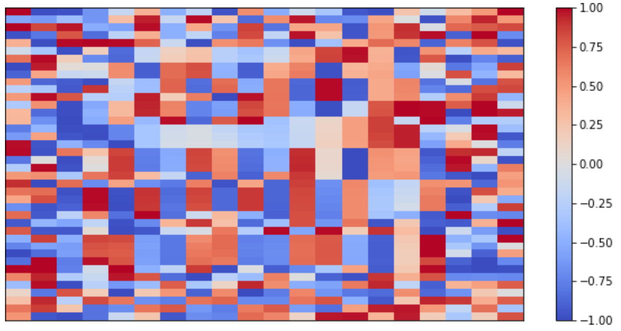


3 phase features on a nonlinear frequency axis

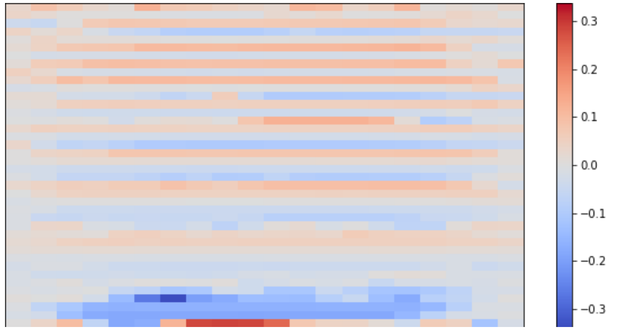
Introduction: Mel-Frequency Feature Representation



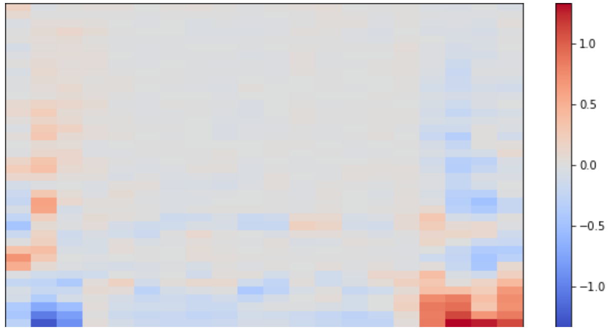
Phase phasor



Instantaneous frequency
(time derivative of the phase)



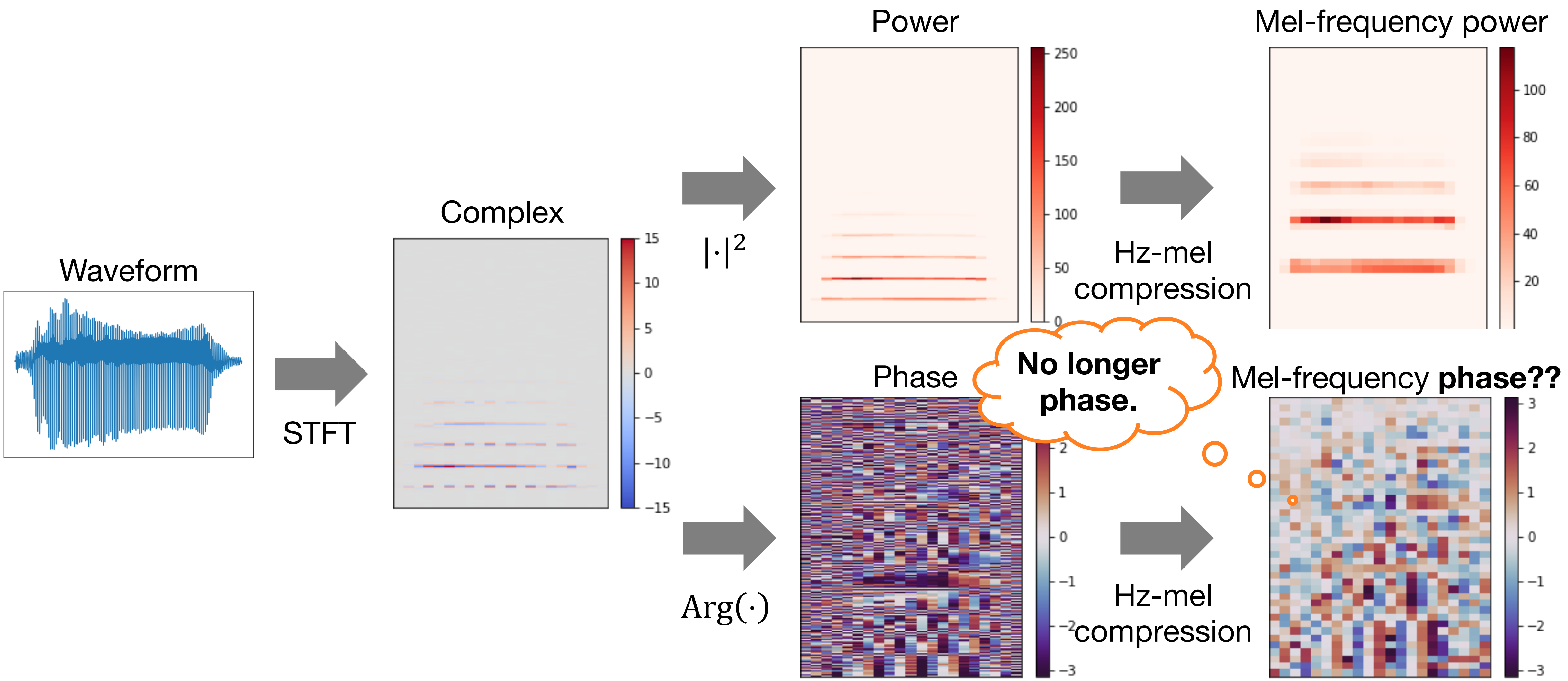
Group delay
(frequency derivative of the phase)



The purpose of this study is to investigate the effectiveness of the phase features for audio classification.

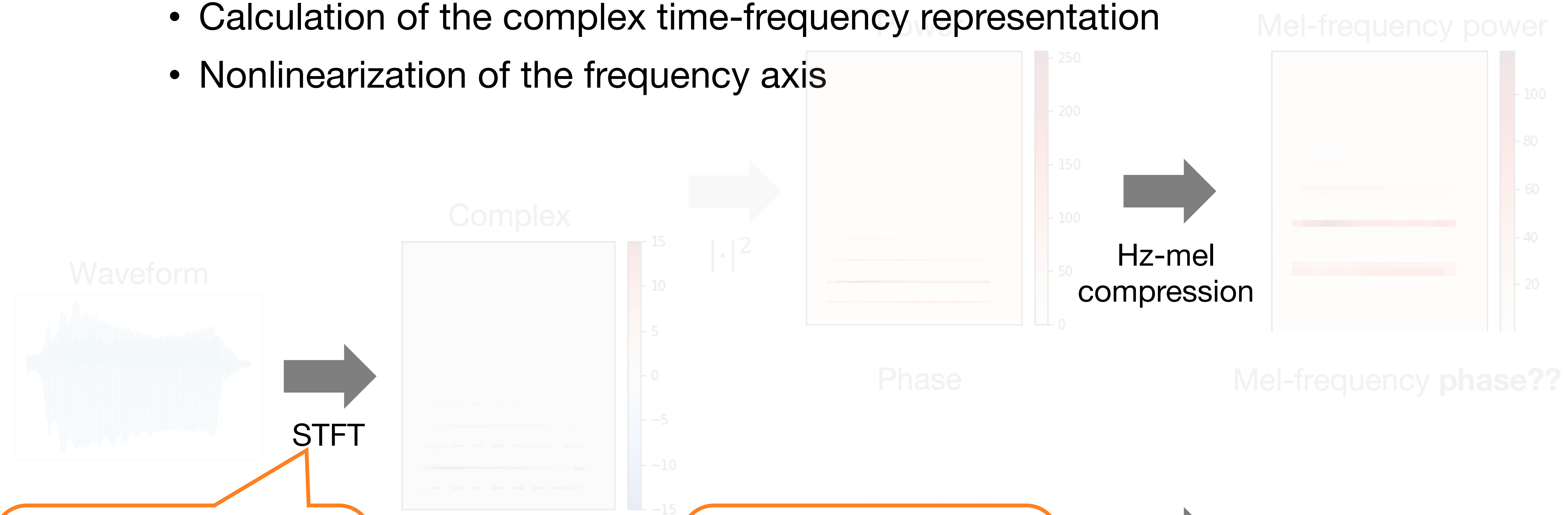
Theory: How Could Be the Phase of Mel Spectrogram?

- The phase of the mel spectrogram is **NOT trivial.**



Theory: How Could Be the Phase of Mel Spectrogram?

- The problem is the separation of the following processes:
  - Calculation of the complex time-frequency representation
  - Nonlinearization of the frequency axis



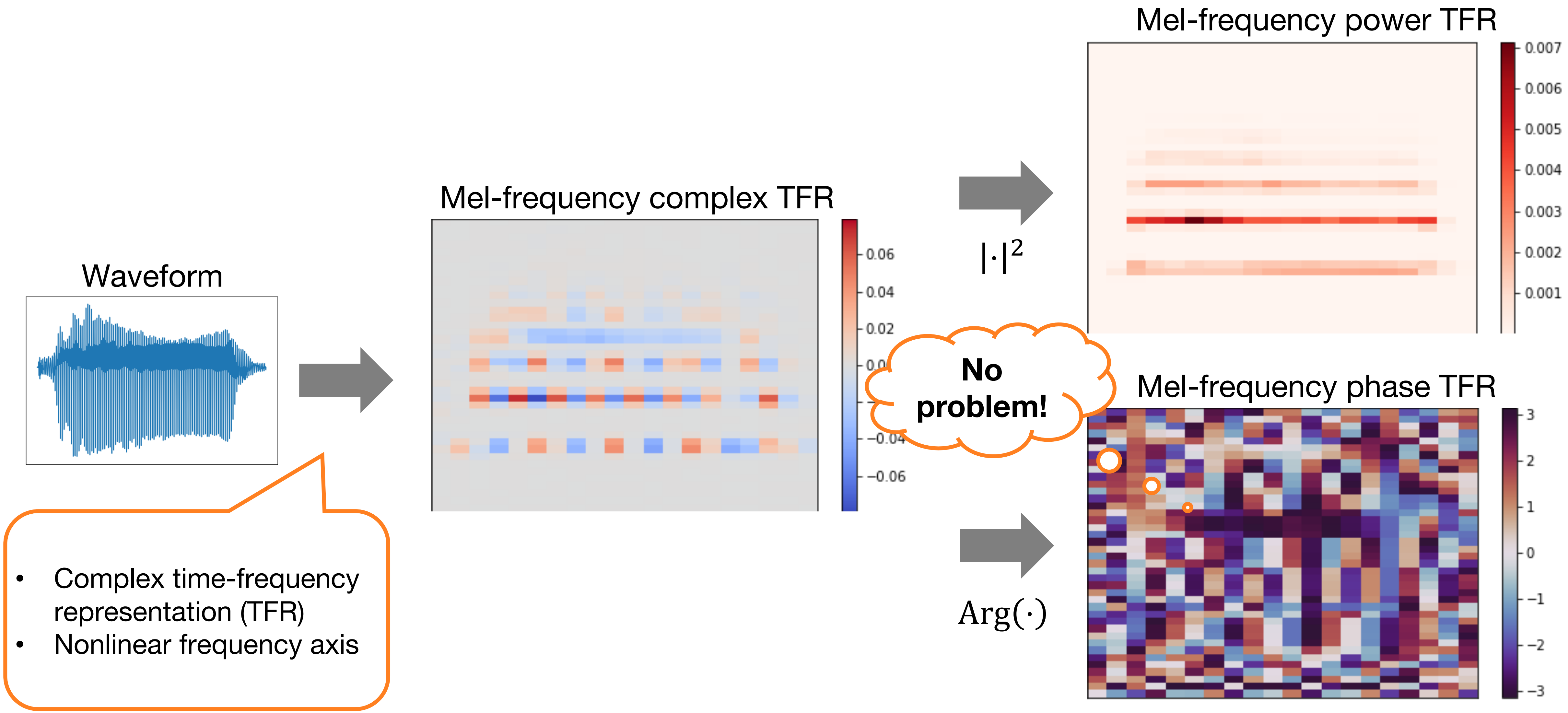
- Complex time-frequency representation (TFR)

Combine

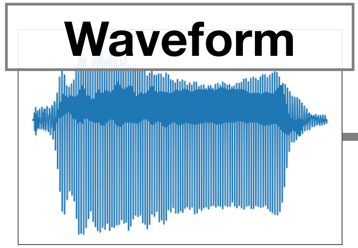
- Nonlinear frequency axis

Hz-mel
compression

Theory: How Could Be the Phase of Mel Spectrogram?



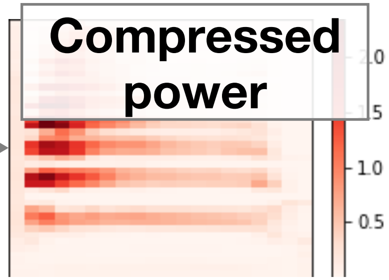
Methods: LEAF-extended

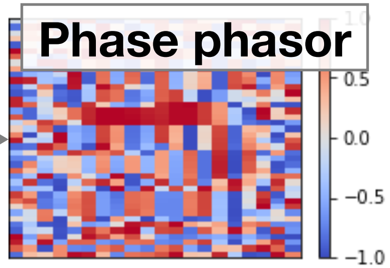


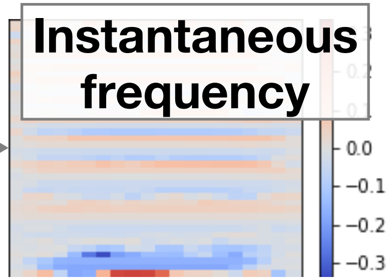
LEAF-extended (LEarnable Audio Frontend - extended)

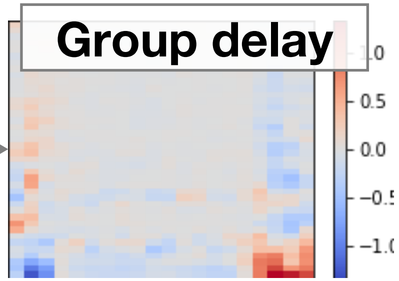
LEAF-extended

- can calculate **power** and **phase** features on a **mel-like nonlinear frequency axis**.
  - Compressed power
  - Phase phasor
  - Instantaneous frequency (time derivative of the phase)
  - Group delay (frequency derivative of the phase)
- is based on LEarnable Audio Frontend (LEAF) [Zeghidour+2021].
  - LEAF only outputs a power feature.
  - LEAF performs comparable with or better than the log-amplitude mel spectrogram [Zeghidour+2021].
- has learnable parameters.

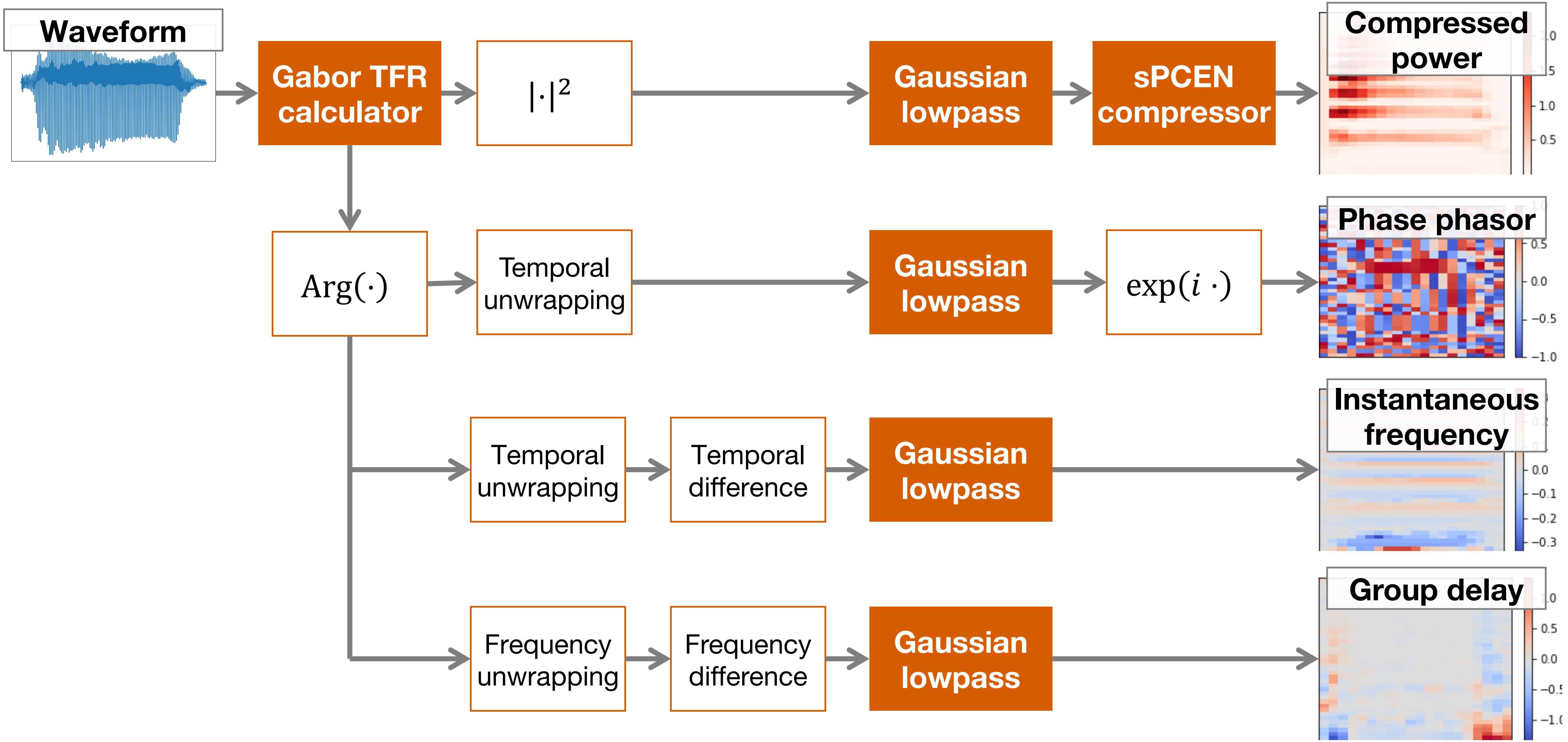




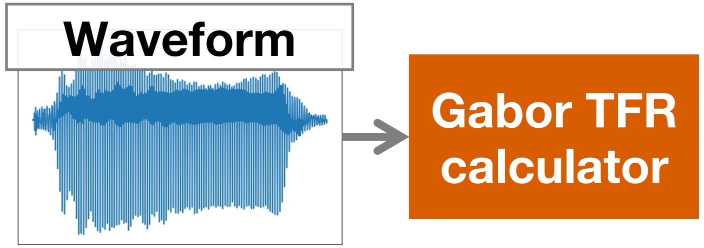


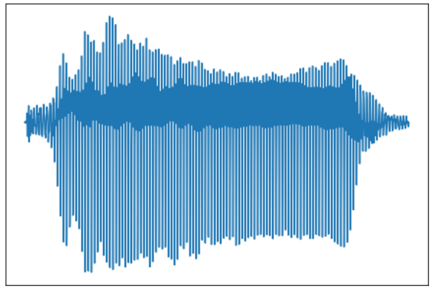


Methods: LEAF-extended



Methods: Gabor Time-Frequency Representation Calculator

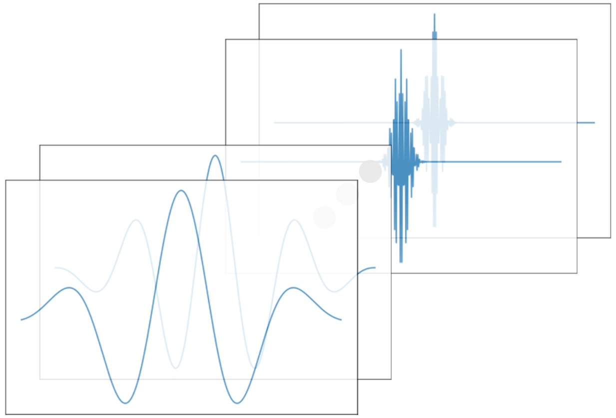




Waveform

$*$

1-D conv



Gabor filterbank

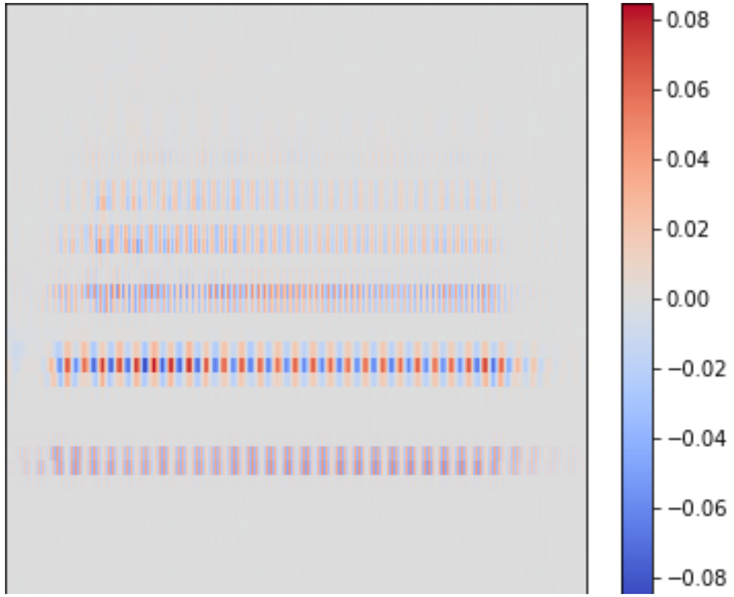
$$\varphi_m(n) = \exp\left(-\frac{n^2}{2\sigma_m^2} + 2\pi i \eta_m n\right)$$

m: filter ID, n: time index,

σ_m: window width (learnable),

η_m: center frequency (learnable)

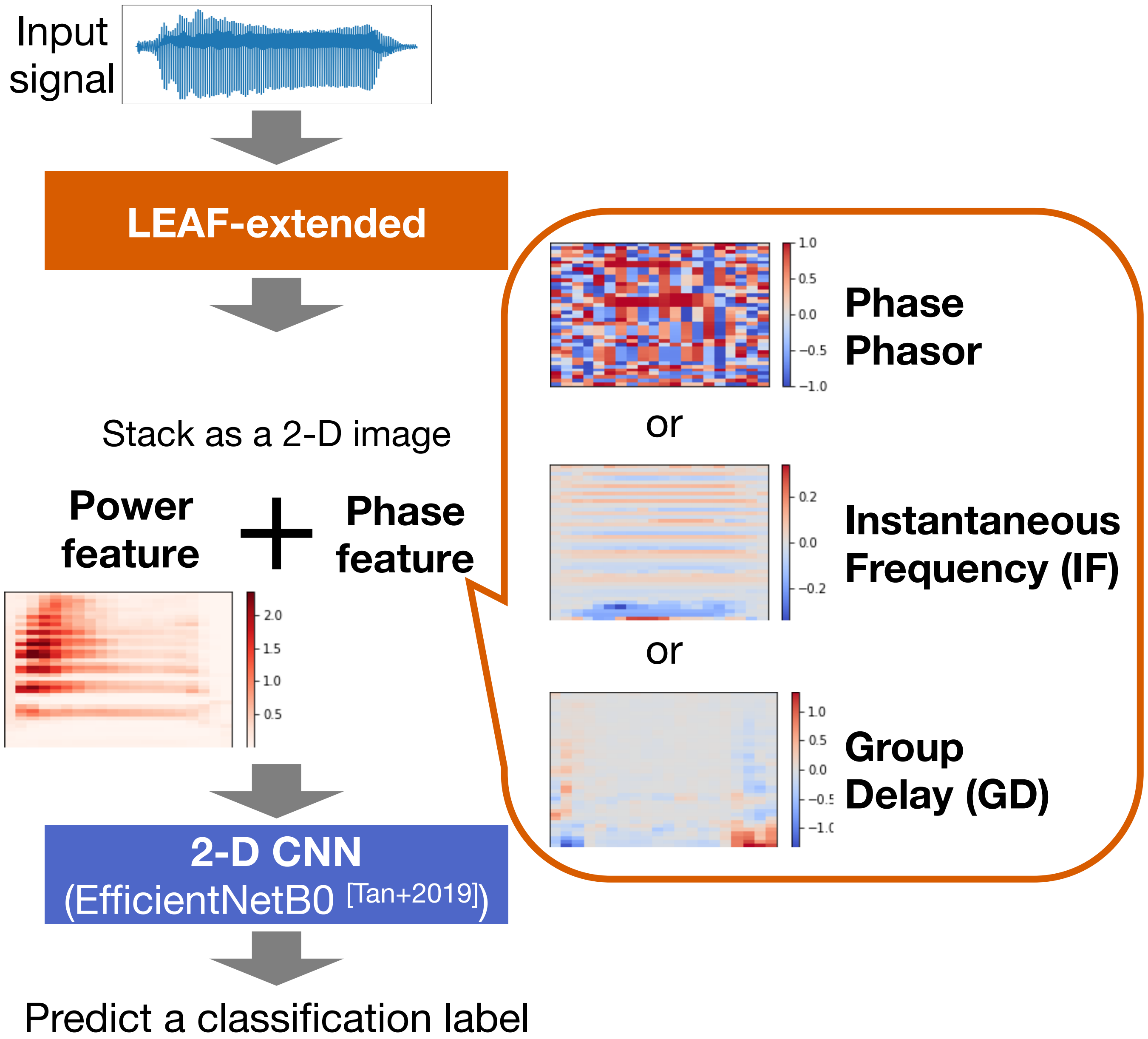
$=$



Mel-like frequency complex TFR

The learnable parameters are initialized so that the frequency response has a similar shape as the mel filterbank.

Experiments: Neural Network for Audio Classification



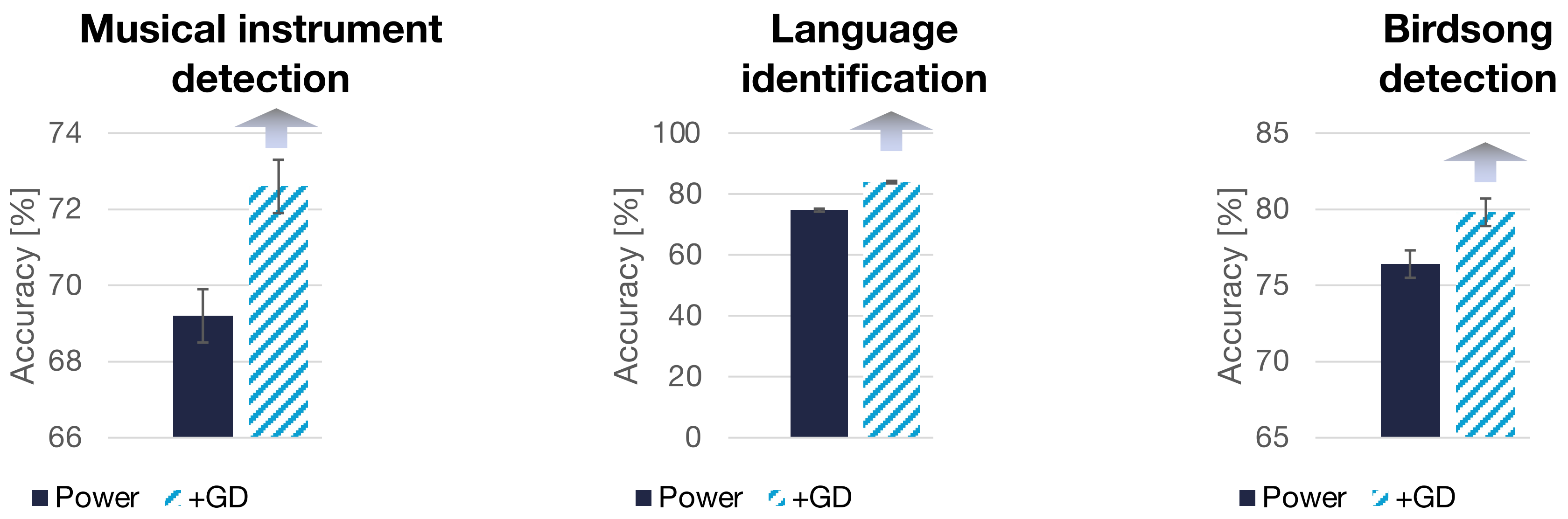
1. LEAF-extended outputs the power and phase features from an input signal.
   - Either one of the phase features is calculated.
2. The features are stacked as a 2-D image.
3. The features are input to a 2-D CNN, and the CNN predicts a classification label.

Experiments: Classification Tasks

- Eight audio classification tasks were performed to investigate the effectiveness of the phase features.

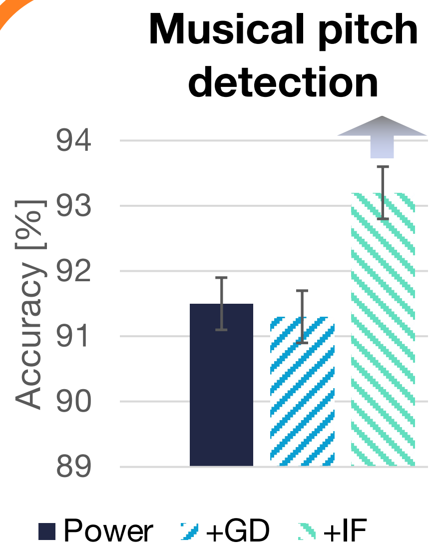
Task	Dataset	Classes	Training samples	Evaluation samples
Musical pitch detection	NSynth [Engel+2017]	112	289,205	16,774
Musical instrument detection	NSynth [Engel+2017]	11	289,205	16,774
Language identification	VoxForge [Revay+2019]	6	148,654	27,764
Birdsong detection	DCASE2018 [Stowell+2018]	2	35,690	12,620
Speaker identification	VoxCeleb [Nagrani+2017]	1,251	128,086	25,430
Acoustic scene classification	TUT [Heittola+2018]	10	6,122	2,518
Keyword spotting	SpeechCommands [Warden2018]	35	84,843	20,986
Emotion recognition	CREMA-D [Cao+2014]	6	5,146	2,296

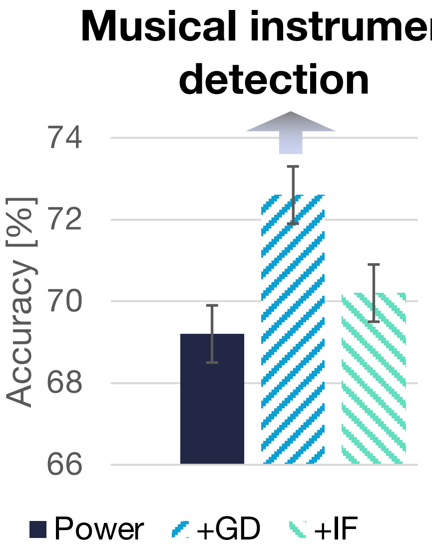
Results and Discussion: Group Delay (GD)

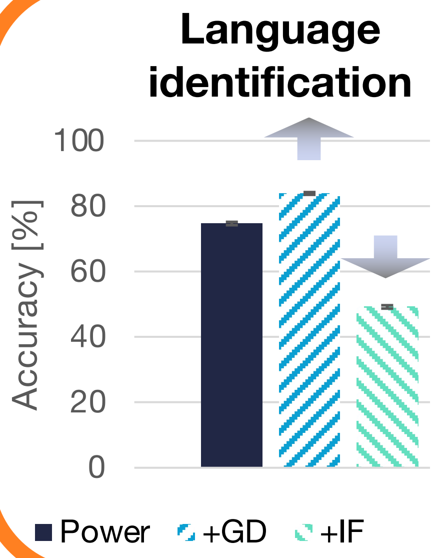


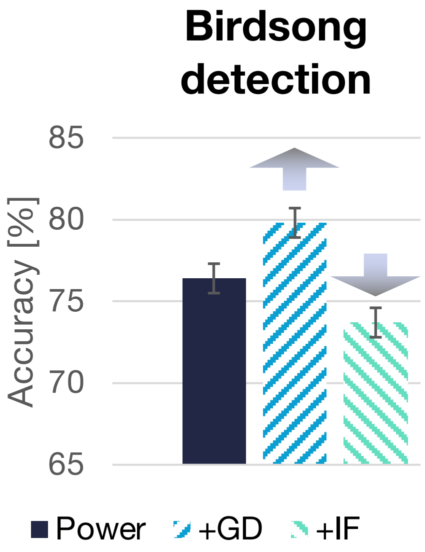
- Compared to using the power alone,
  the performance significantly **improved** by adding GD
  in musical instrument detection, language identification, and birdsong detection.
  - GD has already been applied to
    formant estimation and segmentation of speech [Murthy+2011].
  - GD might include information about timbre and segmentation.

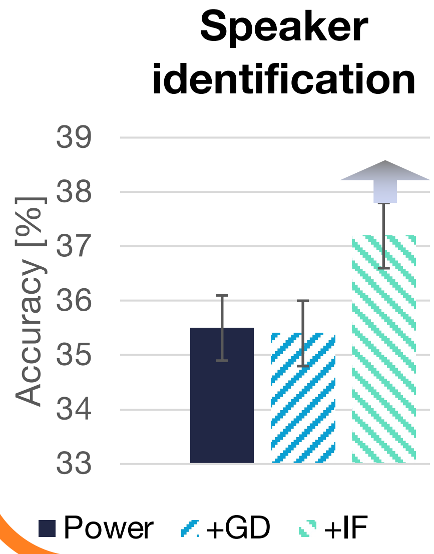
Results and Discussion: Instantaneous Frequency (IF)







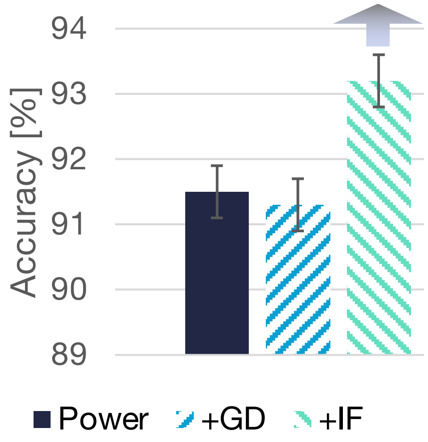




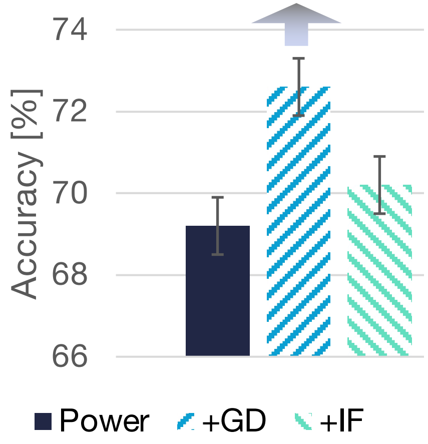
- Compared to using the power alone, the performance significantly **improved** by adding IF in musical pitch detection and speaker identification.
  - IF has already been applied to F0 estimation successfully [Kawahara+2011].

Results and Discussion: Instantaneous Frequency (IF)

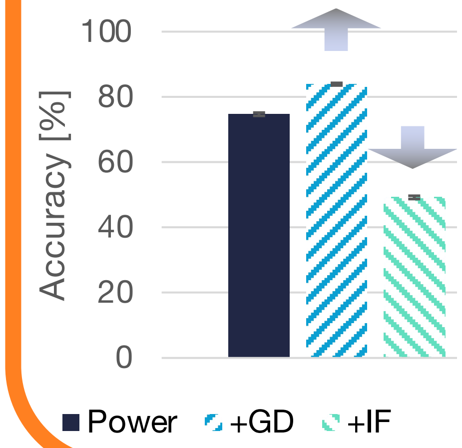
Musical pitch detection



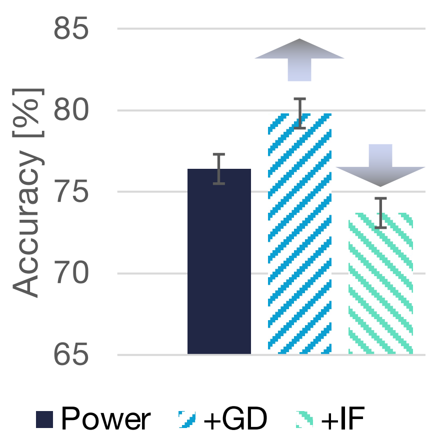
Musical instrument detection



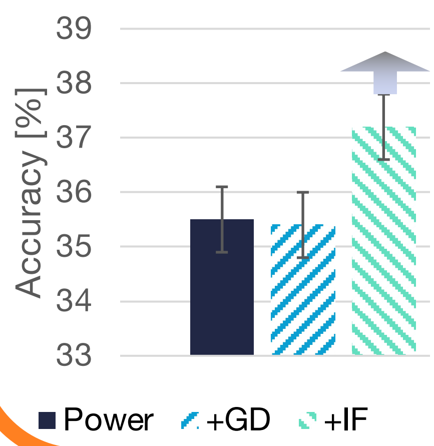
Language identification



Birdsong detection

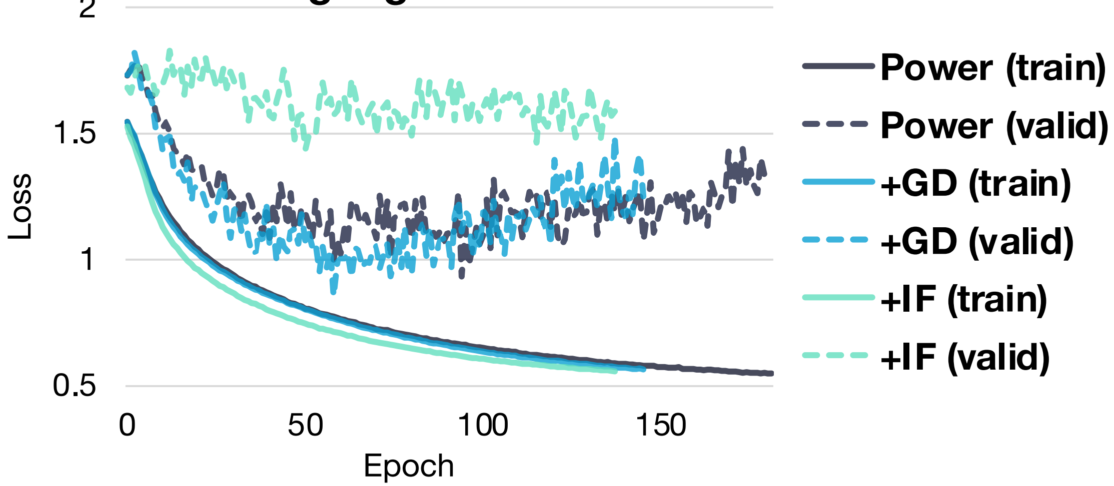


Speaker identification



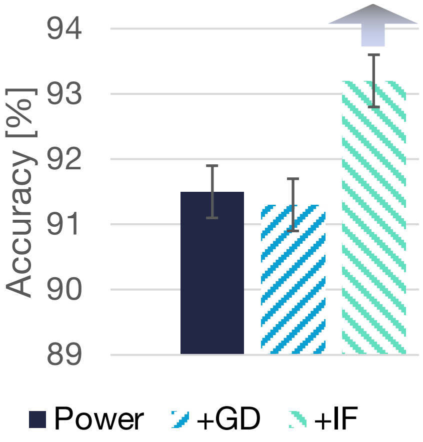
- Compared to using the powe
the performance significantly
in musical pitch detection an
  - IF has already been appl

Learning curves for language identification

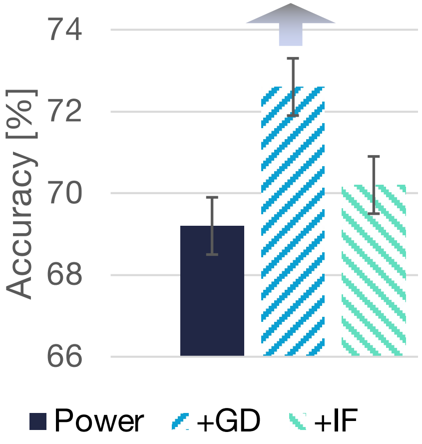


Results and Discussion: Instantaneous Frequency (IF)

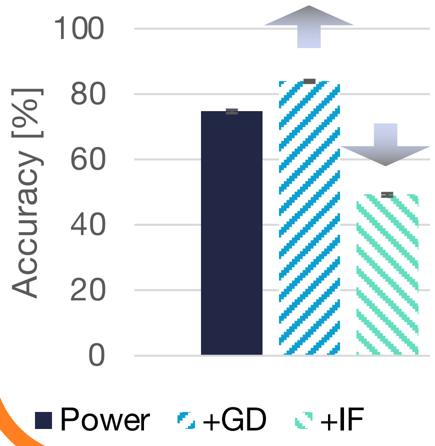
Musical pitch detection



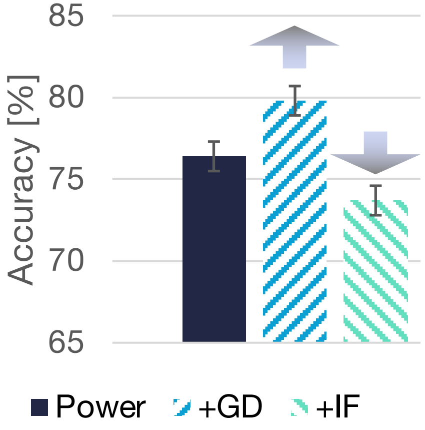
Musical instrument detection



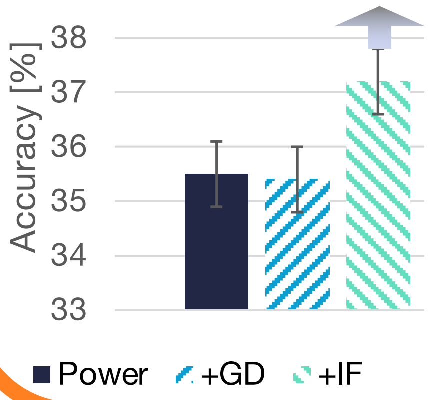
Language identification



Birdsong detection

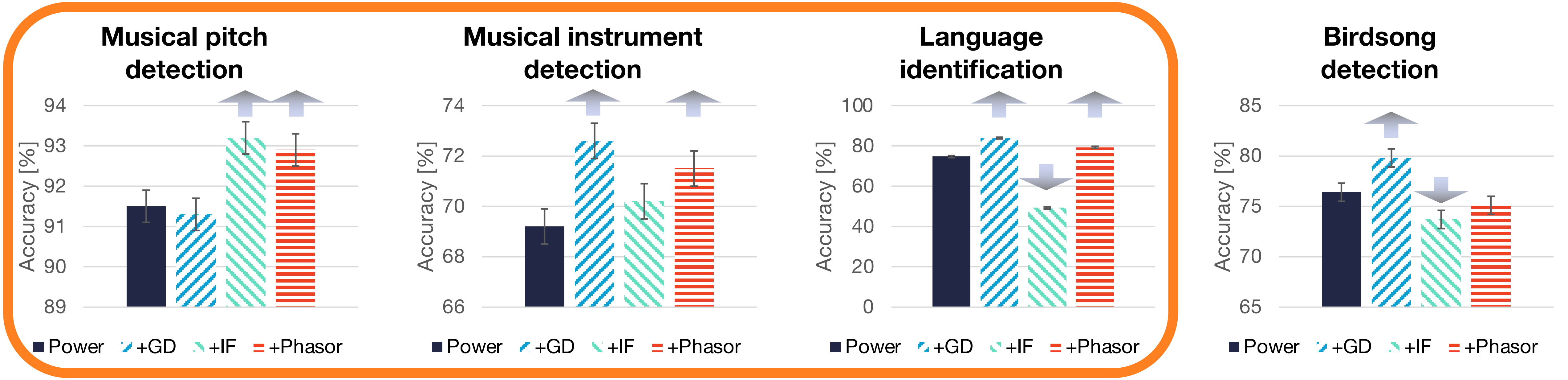


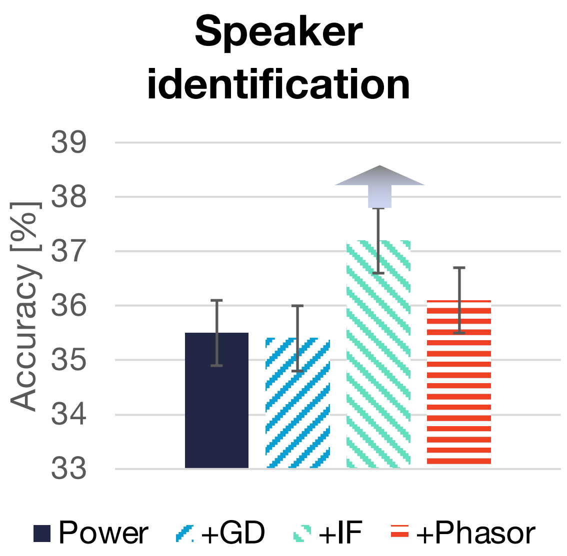
Speaker identification



- Compared to using the power alone, the performance significantly **improved** by adding IF in musical pitch detection and speaker identification.
  - IF has already been applied to F0 estimation successfully [Kawahara+2011].
- The performance significantly ***degraded*** by adding IF in language identification and birdsong detection.
  - The datasets for language identification and birdsong detection contained data from various recording environments (e.g., power line hum).
  - IF might have caused overfitting to the recording environments.

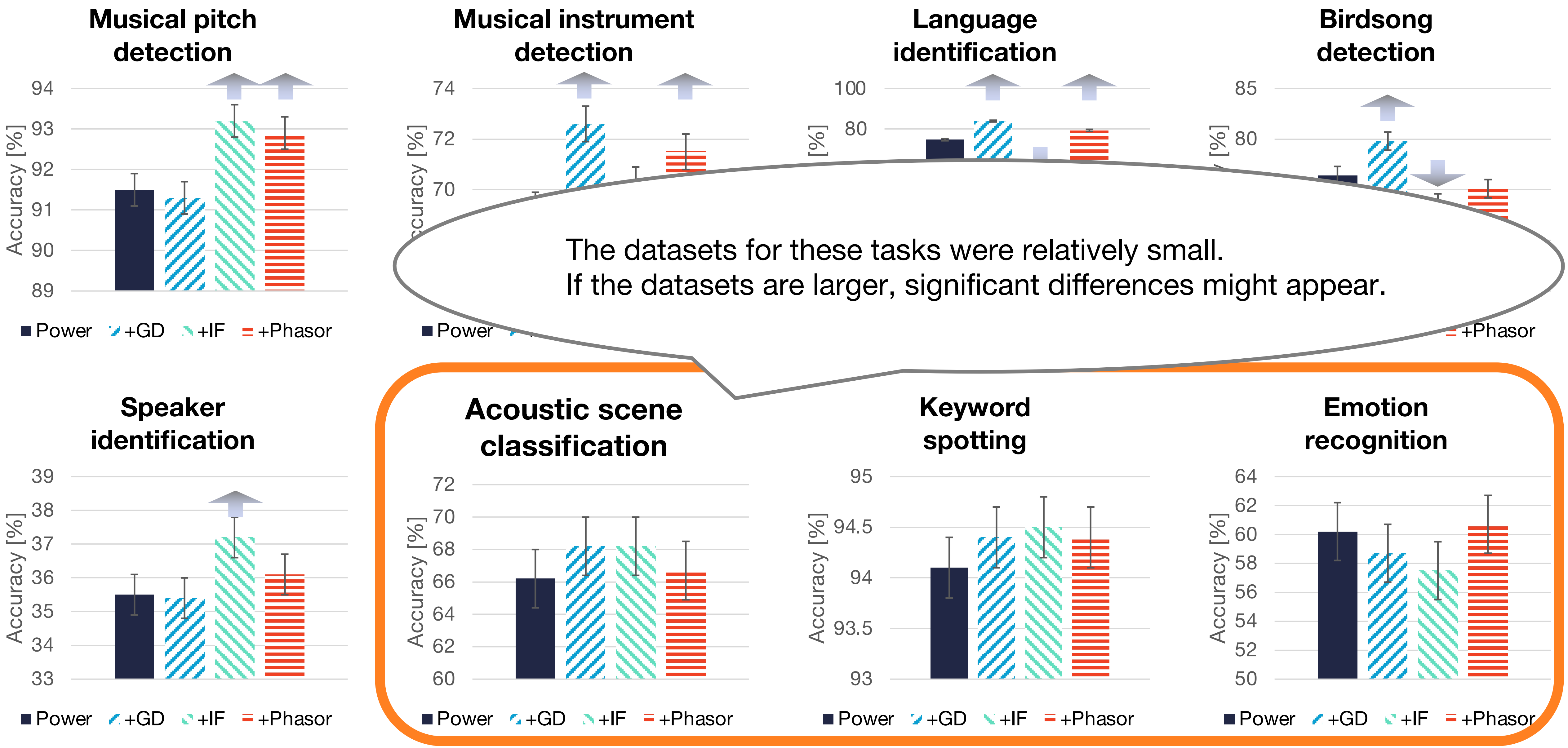
Results and Discussion: Phase Phasor





- Compared to using the power alone,
  the performance significantly **improved** by adding the phase phasor
  in musical pitch detection, musical instrument detection, and language identification.
- For a specific task, if the phase phasor significantly improved performance,
  then the derivatives of the phase (GD or IF)
  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.

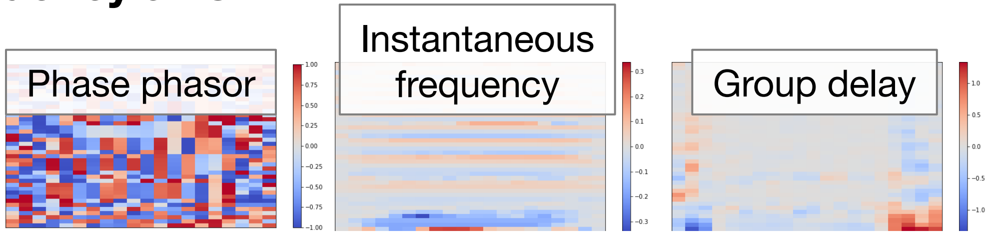
Results and Discussion: Remaining Tasks



The datasets for these tasks were relatively small.
If the datasets are larger, significant differences might appear.

Conclusion

- We investigated the effectiveness of the phase features of a time-frequency representation for audio classification.
- We proposed a learnable audio frontend, LEAF-extended, which can calculate **phase features on a learned nonlinear frequency axis**.
    - Phase phasor
    - Instantaneous frequency (the time derivative of the phase)
    - Group delay (the frequency derivative of the phase)



- The results suggested that **the phase and its derivatives were valuable in some classification tasks**:
    - Musical pitch detection
    - Musical instrument detection
    - Language identification
    - Speaker identification
    - Birdsong detection
- On the other hand, the instantaneous frequency might have caused **overfitting to the recording environments** (e.g., power line hum) in some tasks.
    - Future work should address the impact of recording environments.

Acknowledgment: This work was supported by JST SPRING Grant Number JPMJSP2136.

References

H. Cao *et al.*, “CREMA-D: Crowd-sourced Emotional Multimodal Actors Dataset,” *IEEE Trans Affect Comput.*, vol. 5, no. 4, pp. 377–390, 2014.

J. Engel *et al.*, “Neural audio synthesis of musical notes with wavenet autoencoders,” in *ICML*, 2017, pp. 1068–1077.

T. Heittola *et al.*, “TUT Urban Acoustic Scenes 2018, Development dataset.” https://doi.org/10.5281/zenodo.1228142, 2018.

T. Heittola *et al.*, “Acoustic scene classification in DCASE 2020 challenge: generalization across devices and low complexity solutions,” *arXiv preprint arXiv:2005.14623*, 2020.

S. Hu *et al.*, “Phase-aware music super-resolution using generative adversarial networks,” in *INTERSPEECH*, 2020, pp. 4074–4078.

H. Kawahara *et al.*, “An interference-free representation of instantaneous frequency of periodic signals and its application to F0 extraction,” in *ICASSP*, 2011, pp. 5420–5423.

Y. Luo and N. Mesgarani, “Conv-TasNet: Surpassing ideal Time–Frequency magnitude masking for speech separation,” *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 27, no. 8, pp. 1256–1266, 2019.

H. A. Murthy and B. Yegnanarayana, “Group delay functions and its applications in speech technology,” *Sadhana - Academy Proceedings in Engineering Sciences*, vol. 36, no. 5, pp. 745–782, 2011.

A. Nagrani *et al.*, “Voxceleb: a large-scale speaker identification dataset,” *arXiv preprint arXiv:1706.08612*, 2017.

S. Revay and M. Teschke, “Multiclass Language Identification using Deep Learning on Spectral Images of Audio Signals,” *arXiv preprint arXiv:1905.04348*, 2019.

D. Stowell *et al.*, “Automatic acoustic detection of birds through deep learning: The first Bird Audio Detection challenge,” *Methods Ecol. Evol.*, vol. 10, no. 3, pp. 368–380, Nov. 2018.

M. Tan and Q. Le, “Efficientnet: Rethinking model scaling for convolutional neural networks,” in *ICML*, 2019, pp. 6105–6114.

P. Warden, “Speech Commands: A Dataset for Limited-Vocabulary Speech Recognition,” *arXiv preprint arXiv:1804.03209*, 2018.

N. Zeghidour *et al.*, “LEAF: A Learnable Frontend for Audio Classification,” in *ICLR*, 2021.

Y. Zhang *et al.*, “Pushing the limits of Semi-Supervised learning for automatic speech recognition,” *arXiv preprint arXiv:2010.10504*, 2020.