Zero-Shot Audio Classification with Factored Linear and Nonlinear Acoustic-Semantic Projections

Huang Xie*, Okko Räsänen*†, Tuomas Virtanen*

* Tampere University

† Aalto University

Motivation

- Audio classification with supervised learning techniques:
  - ✓ Requires large amounts of annotated audio data from target classes.
  - ■ Data collection and manual annotation are labor-intensive, time-consuming, and costly.

- Audio classification with limited audio data:
  - ✓ Employs methods such as data augmentation, meta learning, few-shot learning, etc.
  - ■ A certain amount of representative audio data from target classes is still indispensable.

- Audio classification for novel classes:
  - ✓ Requires retraining supervised models.
  - ■ time-consuming, exhaustive parameter tuning, etc.

- An extreme case → no available audio data but only semantic information from target classes

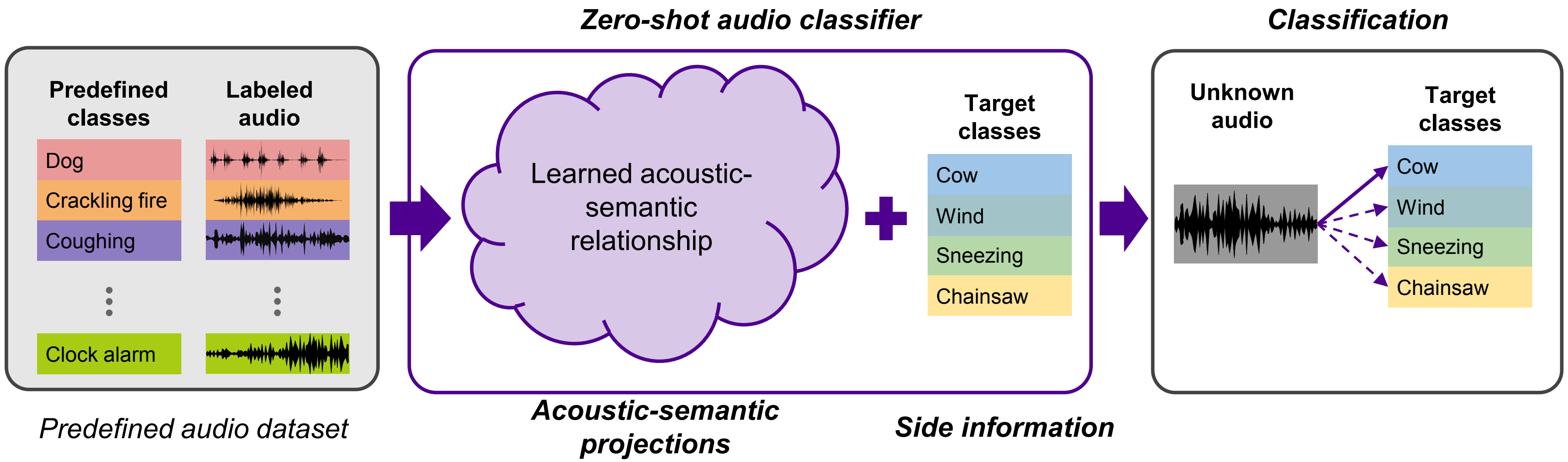
Tampere University

Zero-Shot Audio Classification

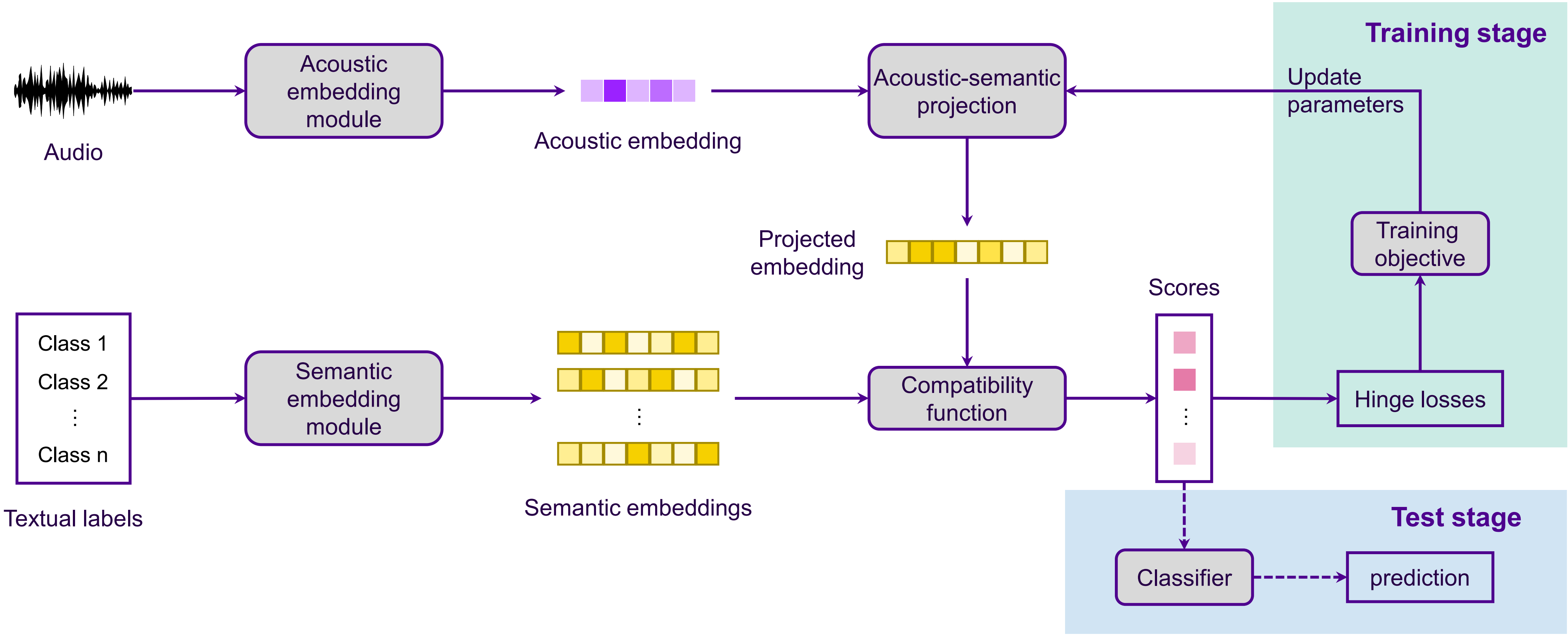
- We tackle the extreme case with zero-shot learning techniques:
  - ✓ Define classes with their **semantic side information**, i.e., class textual labels.
  - ✓ Learn **acoustic-semantic projections** between audio data and textual labels from **predefined training classes**.
  - ✓ Transfer the learned projections to classify audio instances from **target classes** based on their labels.

    ⇒ Target classes are **disjoint** from the predefined training classes.

- The core idea is to model the relationships between audio data and semantic information, i.e., acoustic-semantic projections.



Model-Agnostic Learning Framework



Bilinear Acoustic-Semantic Projection

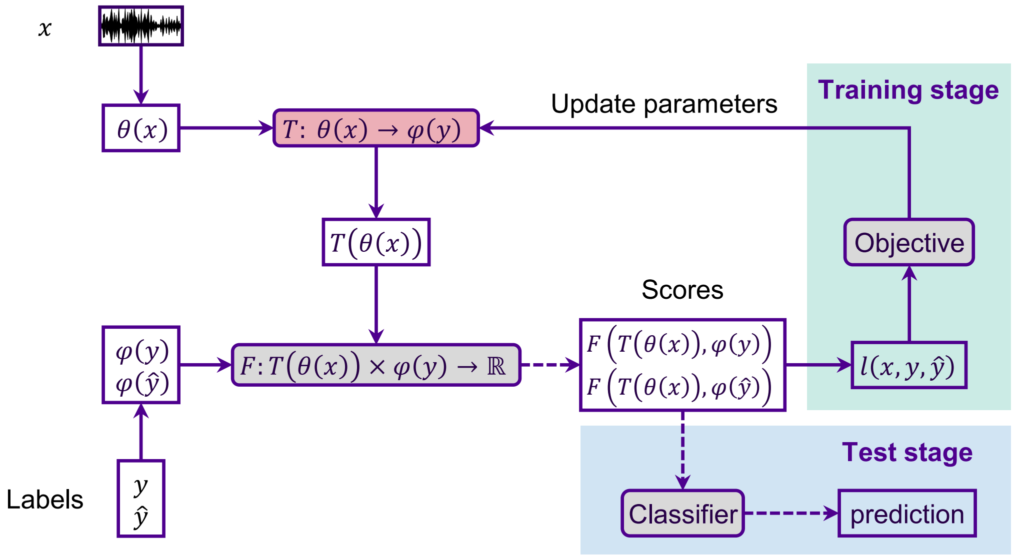
- Given the acoustic embedding $\theta(x)$ of an audio instance x, the semantic embedding $\varphi(y)$ of its reference class y, and $\varphi(\hat{y})$ of class $\hat{y}$.

- Denote the acoustic-semantic projection by T:

  $\Rightarrow$ project $\theta(x)$ onto $\varphi(y)$ such that they are close to each other.

- A simple linear projection with a matrix W:

$$T(\theta(x)) = W'\theta(x)$$



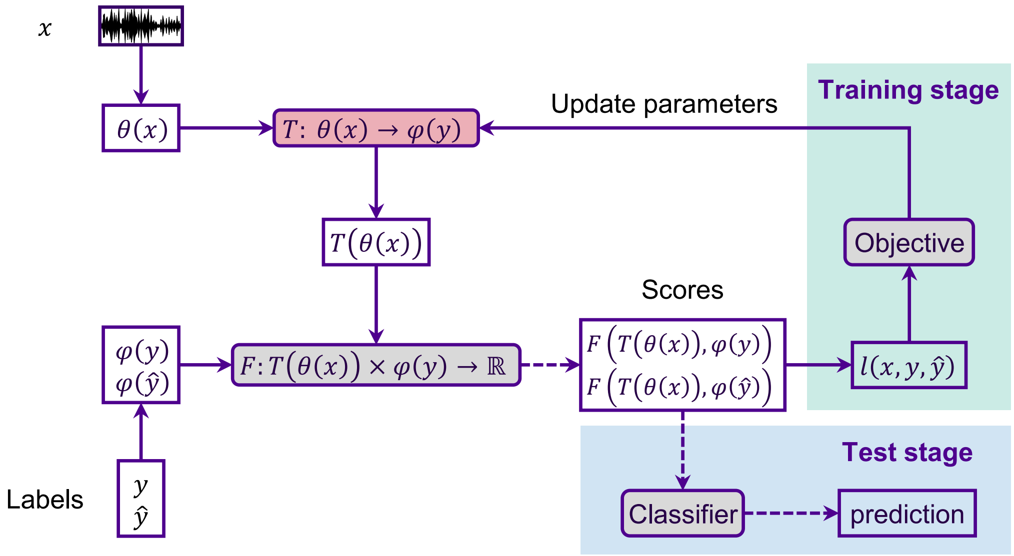
Factored Linear Acoustic-Semantic Projection

- Decompose W into a product of two low-rank matrices U and V.

  $\Rightarrow$ reduce the effective number of learned parameters.

- The factored linear projection:

$$T(\theta(x)) = V'U'\theta(x)$$



Nonlinear Acoustic-Semantic Projections

- Introduce nonlinear activations into factored linear projection.

  ⇒ model possible nonlinearity between acoustic embeddings and semantic embeddings.

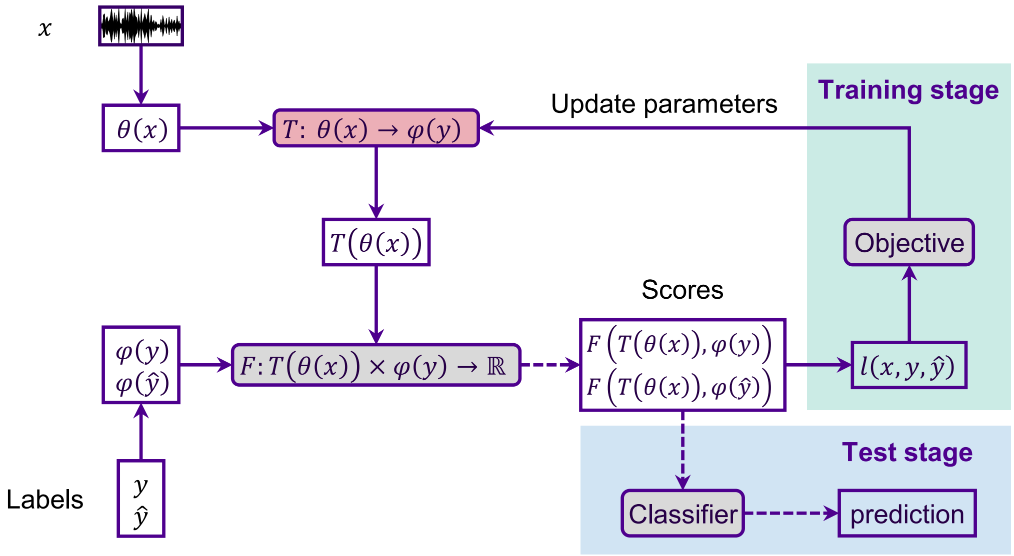
- The nonlinear projection with a nonlinear activation t:

$$T(\theta(x)) = V' t(U' \theta(x))$$

  ⇒ options of t: ReLU, sigmoid, tanh, etc.

- Introduce more projection matrices (e.g., Q) and nonlinear activations t:

$$T(\theta(x)) = V' t\left(Q\, t(U' \theta(x))\right)$$



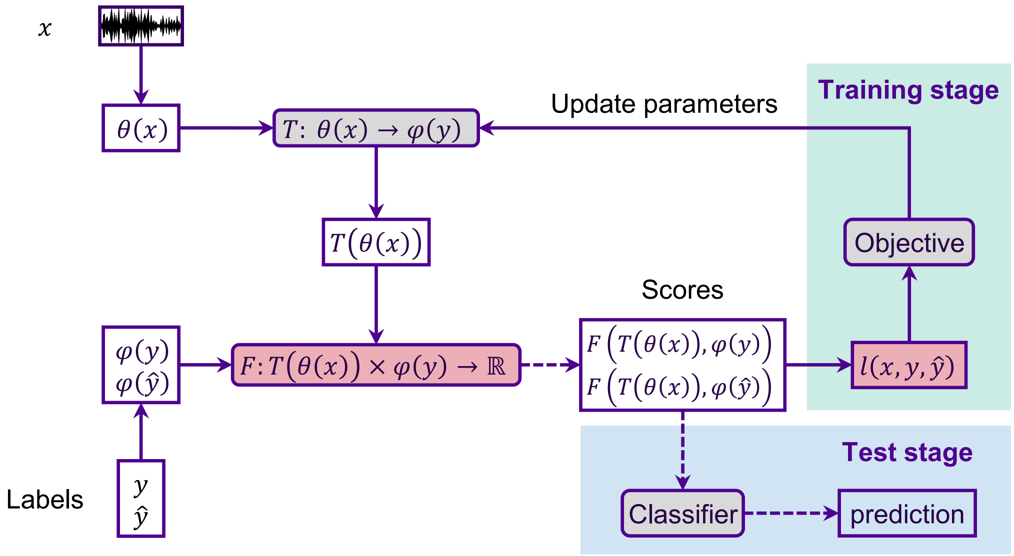
Compatibility Function & Loss

- Choose the dot product as the compatibility function F:

$$F\left(T(\theta(x)),\varphi(y)\right) = T(\theta(x))'\varphi(y)$$

$\Rightarrow$ classify x into a class that has the maximum compatibility.

$\Rightarrow$ other options of F: cosine similarity, etc.



- Define hinge loss $l(x,y,\hat{y})$:

$$l(x,y,\hat{y}) = \max\left(0,\Delta(y,\hat{y}) + F\left(T(\theta(x)),\varphi(\hat{y})\right) - F\left(T(\theta(x)),\varphi(y)\right)\right)$$

$\Rightarrow$ $\Delta(y,\hat{y}) = 0$ if $y = \hat{y}$ and 1 otherwise.

Embedding Modules

- VGGish:
  ⇒ trained from scratch.
  ⇒ extract 128-dimensional acoustic embeddings from audio clips.

- Pre-trained Word2Vec:
  ⇒ generate 300-dimensional semantic embeddings by averaging word vectors in class textual labels.

Evaluation – Dataset

- An unbalanced subset from AudioSet:
  - ✓ 112,774 single-labeled audio clips.
  - ✓ 521 sound classes.
  - ✓ divided into 5 disjoint class folds:
    - ⇒ “Fold0” and “Fold1” for training VGGish.
    - ⇒ “Fold2”, “Fold3”, and “Fold4” for zero-shot classification.

Class Fold	Sound Class	Audio Clips
Fold0	104	23,007
Fold1	104	22,889
Fold2	104	22,762
Fold3	104	22,739
Fold4	105	21,377

Evaluation – Acoustic-Semantic Projections

- Acoustic-semantic projections:
  - ✓ Bilinear projection (baseline)
  - ✓ Factored linear projection
  - ✓ Nonlinear projections:
    - ⇒ two fully-connected layers with ReLU ($FC2_{relu}$), sigmoid ($FC2_{sigmoid}$), tanh activations ($FC2_{tanh}$).
    - ⇒ three fully-connected layers with tanh activations ($FC3_{tanh}$).

- To prevent randomness, each projection is evaluated twenty times with random initialization.

Evaluation – Results

- With $FC2_{sigmoid}$ and $FC2_{tanh}$,
  ⇒ Capture nonlinearity between acoustic and semantic embeddings.
  ⇒ Improve zero-shot performance.
- With $FC3_{tanh}$,
  ⇒ no explicit benefit with more parameters and nonlinear activations.

Acoustic-Semantic Projection		TOP-1 (%) avg $\pm$ std
Bilinear (baseline)		5.7 ± 1.1
Factored Linear		6.3 ± 0.8
Nonlinear	$FC2_{relu}$	5.5 ± 0.9
	$FC2_{sigmoid}$	$\mathbf{7.0 \pm 0.5}$
	$FC2_{tanh}$	$\mathbf{7.2 \pm 0.6}$
	$FC3_{tanh}$	6.0 ± 0.6

Conclusions

- We investigated acoustic-semantic projections for zero-shot learning in audio classification.

  ⇒ Factored linear projection is developed by applying matrix decomposition to a bilinear model.

  ⇒ Nonlinear activations are used to capture nonlinearity between acoustic and semantic embeddings.

  ⇒ A model-agnostic learning framework is used to study the effectiveness of acoustic-semantic projections.