Diffusion Networks for Audio Zero-Shot Learning Further Information

Class index	**38**	**29**	**35**	**3**	**40**	**2**	**27**	**46**	**31**	**48**
38	1	**9**	0	3	0	3	**9**	0	**14**	1
29	1	**18**	0	0	1	0	**18**	0	1	1
35	1	0	6	0	**30**	1	1	1	0	0
3	0	0	5	0	**29**	5	1	0	0	0
40	0	0	6	0	**32**	0	0	1	0	1
2	0	6	3	0	3	**5**	**21**	0	0	2
27	2	**34**	0	1	0	0	1	0	2	0
46	0	0	2	0	**28**	2	0	**8**	0	0
31	1	5	0	0	2	0	6	0	**25**	1
48	4	6	0	0	2	0	**8**	1	**15**	4

TABLE I

Confusion Matrix for fold 0 of the ESC-50 dataset. Rows represent the true class, and columns represent the predictions. Many predictions are clustered around class 40, with smaller hubs on classes 29, 27 and 31.

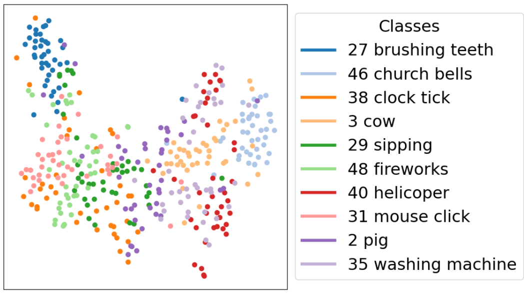

Fig. 1. ESC-50 fold 0 audio embeddings with t-SNE dimensionality reduction. Class 40 (helicopter), 46 (church bells), 35 (washing machine) and 3 (cow) are all concentrated to the right side of the scatterplot. The confusion matrix in Table I shows that all of these classes are classified as helicopter a majority of the time.

Class index	**39**	**36**	**42**	**13**	**32**	**22**	**19**	**49**	**26**	**21**
39	**10**	0	0	0	**9**	0	2	3	0	**16**
36	0	**24**	3	4	0	0	6	1	0	2
42	0	0	**32**	3	0	0	0	2	3	0
13	0	4	7	**14**	1	0	5	0	1	**8**
32	2	0	0	0	**23**	0	2	**13**	0	0
22	**14**	1	0	1	3	0	**14**	6	1	0
19	0	5	**10**	0	0	0	**22**	0	0	3
49	1	3	0	1	0	3	0	**11**	**8**	**13**
26	4	0	2	4	0	0	0	4	**15**	**11**
21	2	0	0	1	1	0	0	0	**8**	**28**

TABLE II

Confusion matrix for fold 1 of the ESC-50 dataset. Rows represent the true class, and columns represent the predictions. There are small hubs on classes 42, 19, 26 and 21 and all classes are at least partially correctly identified, except for class 22 which is never correctly classified.

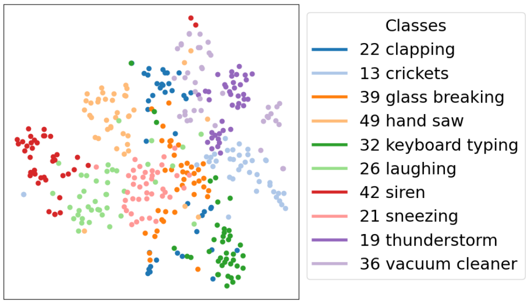

Fig. 2. ESC-50 fold 1 audio embeddings with t-SNE dimensionality reduction. Fold 1 exhibits more defined class clusters than fold 0. Class 22 (clapping, dark blue) has a greater spread, and is the poorest classified class as shown in Table II.

Class index	**10**	**45**	**4**	**14**	**17**	**30**	**41**	**33**	**24**	**23**
10	5	6	0	0	1	4	**20**	3	0	1
45	**11**	1	0	0	0	0	**28**	0	0	0
4	2	2	**15**	**10**	5	0	2	0	2	2
14	0	2	**24**	6	0	1	1	4	2	0
17	1	6	1	4	**14**	1	0	0	1	**12**
30	**22**	0	0	0	5	2	1	0	9	1
41	0	2	0	0	1	0	**27**	6	1	3
33	2	4	**8**	2	5	1	7	2	5	4
24	0	0	1	0	0	0	0	0	**38**	1
23	2	4	1	1	0	2	3	2	**21**	4

TABLE III

Confusion matrix for fold 2 of the ESC-50 dataset. Rows represent the true class, and columns represent the predictions. The matrix is sparse with hubs around classes 4 and 41. While there is some spread, classes 45, 30 and 33 have low amounts of predictions.

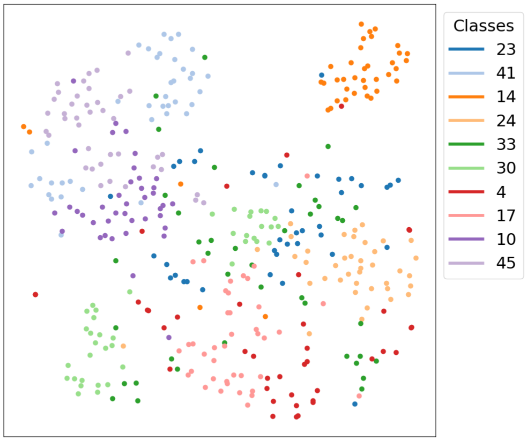

Fig. 3. ESC-50 fold 2 audio embeddings with t-SNE dimensionality reduction. The confusion matrix shows a large hub around class 41 with classes 10, 45 and 41. The t-SNE diagram shows these classes in the top left corner with some overlap and close proximity.

Class index	**6**	**1**	**28**	**18**	**25**	**34**	**20**	**47**	**7**	**44**
6	2	2	3	0	4	0	**28**	0	1	0
1	**4**	8	**12**	0	0	0	**14**	0	2	0
28	2	3	5	1	0	0	0	4	**24**	1
18	3	0	0	**19**	0	0	4	3	**11**	0
25	0	0	0	**11**	6	**14**	0	0	4	5
34	0	0	1	**21**	4	9	0	0	5	0
20	**11**	7	4	0	0	0	**18**	0	0	0
47	0	1	0	0	0	0	0	**29**	7	3
7	1	7	0	2	3	0	1	**13**	6	7
44	1	1	1	8	0	1	2	**11**	**13**	2

TABLE IV
CONFUSION MATRIX FOR FOLD 3 OF THE ESC-50 DATASET. ROWS REPRESENT THE TRUE CLASS, AND COLUMNS REPRESENT THE PREDICTIONS. THERE ARE FEW PREDICTIONS FOR CLASSES 1, 25 AND 44, WHILE CLASSES 18, 20 AND 47 APPEAR AS HUBS.

Class index	**37**	**11**	**9**	**8**	**0**	**15**	**5**	**43**	**16**	**12**
37	**13**	4	0	2	1	1	**18**	1	0	0
11	0	4	0	0	0	**10**	0	7	**19**	0
9	1	4	0	**15**	2	4	**8**	4	2	0
8	4	2	0	**10**	3	0	4	**17**	0	0
0	0	0	1	**28**	2	6	1	0	2	0
15	**13**	2	4	2	2	**12**	4	1	0	0
5	**6**	0	5	**15**	5	0	**9**	0	0	0
43	**8**	1	0	**12**	2	0	4	**12**	1	0
16	2	4	1	2	0	0	1	**18**	**11**	1
12	4	5	0	0	0	6	2	1	3	**19**

TABLE V
CONFUSION MATRIX FOR THE TEST FOLD OF THE ESC-50 DATASET. ROWS REPRESENT THE TRUE CLASS, AND COLUMNS REPRESENT THE PREDICTIONS. CLASSES 11, 9 AND 0 HAVE MINIMAL PREDICTIONS. OTHER CLASSES FORM SMALL HUBS, WITH A LARGE HUB AROUND CLASS 8. CLASS 12 IS UNIQUELY PREDICTED CORRECTLY ALMOST HALF OF THE TIME AND THERE IS ONLY ONE INSTANCE OF ANOTHER CLASS MISCLASSIFIED AS CLASS 12.

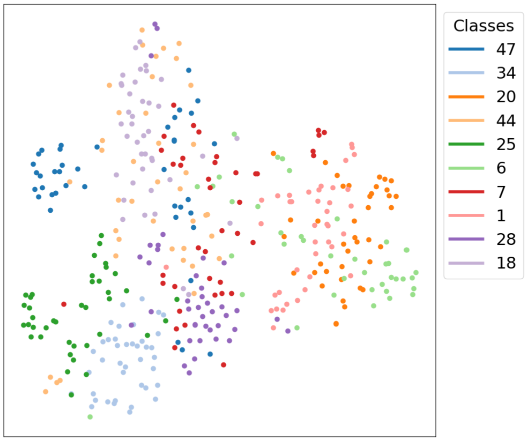

Fig. 4. ESC-50 fold 3 audio embeddings with t-SNE dimensionality reduction. The differences in behaviours of hubs is noticeable in this diagram. The hubs on classes 18 and 47 both cover a large area but are generally close in proximity and overlap. The hub around class 20 is more defined, with classes 20, 1 and 6 almost solely occupying the right hand side of the diagram.

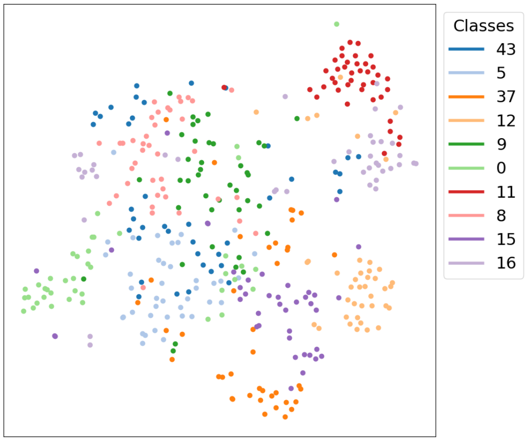

Fig. 5. ESC-50 fold 3 audio embeddings with t-SNE dimensionality reduction. An interesting observation from the confusion matrix is the performance of class 12. This class forms a compact cluster with minimal overlap with other classes. The large cluster around class 18 is very spread and contains overlap with many classes. The classes not in this hub are clustered on the outside of the plot to the right, and have less spread.

Class index	**22**	**18**	**12**	**9**	**6**	**13**	**8**
22	**43**	0	3	4	10	1	14
18	0	0	0	**49**	11	**15**	0
12	**30**	0	**18**	**16**	6	1	4
9	**24**	**44**	0	0	0	6	1
6	3	2	**15**	**24**	11	10	10
13	0	0	0	1	0	**71**	3
8	5	0	8	0	6	0	**56**

TABLE VI

Confusion matrix for the validation fold of the FSC22 dataset. Rows represent the true class, and columns represent the predictions. Compared to ESC-50, there are less classes and more samples per class. The matrix is sparse, similarly to the confusion matrices from ESC-50. Class 22, 13 and 8 all have high correct classifications. Classes 22 and 9 appear to be hubs.

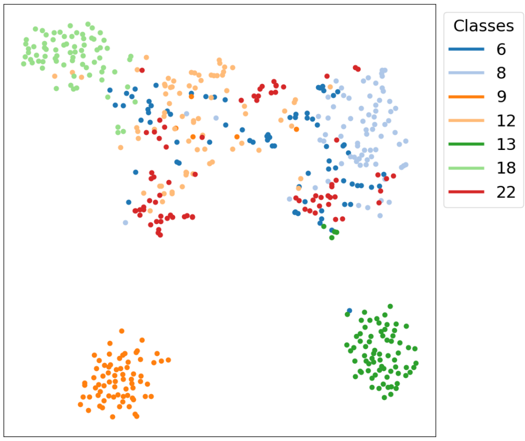

Fig. 6. FSC22 validation fold audio embeddings with t-SNE dimensionality reduction. This t-SNE plot has a different appearance to those in ESC-50, and may indicate that the small number of classes is difficult to draw conclusions with. Classes 8 and 13 have somewhat defined clusters, which may cause their higher classification accuracy, however the isolated clustering of class 9 does not follow the same pattern. Class 9 is a hub, and while classes 18 and 12 are closer than some other classes in the plot, there is a large distance between them. Class 6, which is poorly classified, has a large spread.

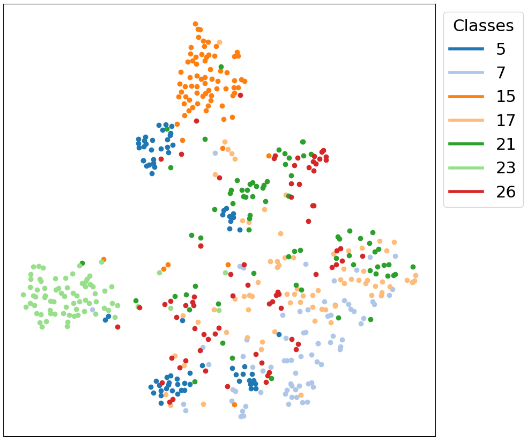

Fig. 7. FSC22 test fold audio embeddings with t-SNE dimensionality reduction. From the confusion matrix, the largest hub exist around classes 26 and 7. Both of the classes have a large spread over other classes, motivating the creation of a hub. Classes 17 and 5 are poorly predicted and have a large, sparse spread.

Class index	**26**	**15**	**7**	**21**	**23**	**17**	**5**
26	**21**	8	**26**	7	8	4	1
15	0	**49**	0	**23**	1	2	0
7	**23**	4	**47**	0	1	0	0
21	**41**	6	8	**16**	3	1	0
23	9	**43**	2	0	**17**	0	4
17	**41**	0	**25**	6	3	0	0
5	7	11	**29**	8	**16**	2	2

TABLE VII

Confusion matrix for the test fold of the FSC22 dataset. Rows represent the true class, and columns represent the predictions. Large hubs exist around classes 26 and 7, and a lack of predictions exist on classes 17 and 5.

Class	**Synonyms**
dog	canine, bark, woof, yap, call, animal, puppy
rooster	cockerel, call, animal
pig	hog, sow, swine, squeal, oink, grunt, call, animal
cow	moo, call, bull, oxen, animal
frog	toad, croak, call, animal
cat	meow, mew, purr, hiss, chirp, kitten, feline, call, animal
hen	cluck, chicken, animal, call
insects (flying)	buzz, hum, bug
sheep	bleat, animal, call, lamb
crow	squawk, screech, caw, bird, call, cry, animal
rain	drizzle, wet, sprinkle, shower, water, nature
sea waves	water, swell, tide, ocean, surf, nature
crackling fire	hissing, sizzling, flame, bonfire, campfire, nature
crickets	insects, insect, bug, cicada, call
chirping birds	animal, call, song, tweet, chirp, twitter, trill, warble, chatter, cheep
water drops	splash, droplet, drip
wind	nature, gust, gale, blow, breeze, howl
pouring water	slosh, gargle, splash, splosh
toilet flush	water, flow, wash
thunderstorm	thunder, storm, nature, lightning
crying baby	cry, human, whine, infant, child, wail, bawl, sob, scream, call
sneezing	sneeze
clapping	clap, applause, applaud, praise
breathing	breath, breathe, gasp, exhale
coughing	cough, hack
footsteps	walking, walk, pace, step, gait, march
laughing	cackle, laugh, chuckle, giggle, funny
brushing teeth	scrape, rub, brush
snoring	snore, sleep, snore, snort, wheeze, breath
drinking, sipping	gulp, gargle, drink, sip, breath
door knock	wood, tap, bang, thump
mouse click	computer, tap
keyboard typing	tap, mechanical, computer
door, wood creaks	squeak, creak, screech, scrape
can opening	hiss, fizz, air
washing machine	electrical, hum, thump, noise, loud
vacuum cleaner	electrical, noise, loud
clock alarm	signal, buzzer, alert, ring, beep
clock tick	tock, click, clack, beat, tap, ticking
glass breaking	crunch, crack, smash, clink, break, noise
helicopter	chopping, engine, blades, whirring, swish, chopper, electrical, noise, vehicle, loud
chainsaw	saw, electrical, noise, tool, loud
siren	alarm, alert, bell, horn, noise, loud
car horn	vehicle, noise, blast, loud, honk
engine	rumble, vehicle, chug, revving, car, drive
train	clack, horn, clatter, vehicle, squeal, rattle
church bells	tintinnabulation, ring, chime, bell
airplane	plane, motor, engine, hum, loud, noise
fireworks	burst, bang, firecracker
hand saw	squeak, sawing, cut, hack, tool

TABLE VIII
Synonyms used for each class in ESC-50 to enhance the class embeddings.

Class	**Synonyms**
fire	crackling, hissing, sizzling, flame, bonfire, campfire, nature
rain	drizzle, wet, sprinkle, shower, water, nature
thunderstorm	thunder, storm, nature, lightning
water drops	splash, droplet, drip
wind	nature, gust, gale, blow, breeze, howl
silence	quiet, silent, soft, nature
tree falling	crackling, wood, nature, crash
helicopter	chopping, engine, blades, whirring, swish, chopper, electrical, noise, vehicle, loud
vehicle engine	rumble, chug, revving, car, drive
axe	chop, cutting, wood, tool
chainsaw	saw, electrical, noise, tool, loud
generator	hum, electrical, machine
hand saw	squeak, sawing, cut, hack, tool
fireworks	burst, bang, firecracker
gunshot	gun, firearm, weapon, shot
wood chop	breaking, splintering, crack
whistling	whistle, high, pitch
speaking	talking, speech, conversation
footsteps	walking, walk, pace, step, gait, march
clapping	clap, applause, applaud, praise
insect	flying, buzz, hum, bug
frog	toad, croak, call, animal
bird chirping	animal, call, song, tweet, chirp, twitter, trill, warble, chatter, cheep
wing flapping	flap, bird, animal
lion	roar, growl, call, animal
wolf howl	canine, call, animal
squirrel	call, animal, chatter, chirp, bark, whistle

TABLE IX
SYNONYMS USED FOR EACH CLASS IN FSC22 TO ENHANCE THE CLASS EMBEDDINGS.