Audio Language Classifier

Kirsten Regier

December 14, 2020

Why Audio Classifiers?

Goal: Using the VGGish model as a feature extractor, build and train a classifier to identify speaker characteristics (gender or native language) from audio files.

Automatic speech recognition (ASR) and voice assistants are expanding.

ASR systems don't work well with non-standard or non-native accents [Sheng & Edmund, 2017].

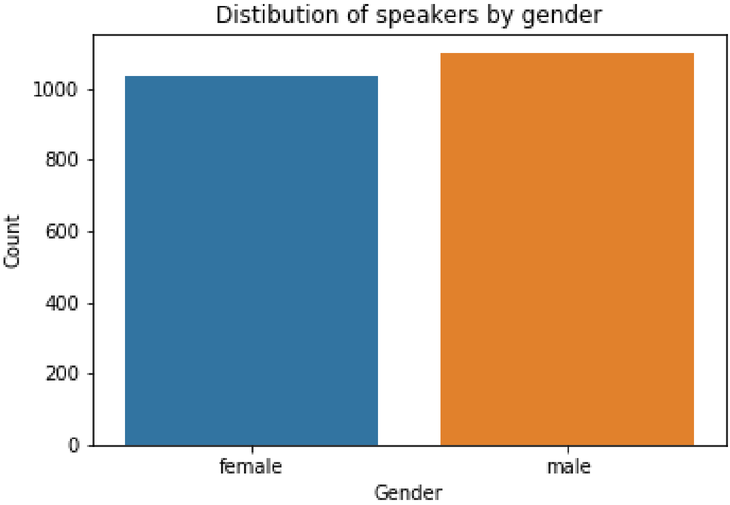
Better speech technology may improve customer experiences and customer service.

The Data

Speech Accent Archive [Weinberger, 2013]

- 2000+ recordings of speech samples.
- Speakers read a fixed passage in English.
- Demographic information about speakers: age, sex, birthplace, country, native language.
  - Speakers from 177 countries.
  - Native speakers of 199 native languages.
  - 78 languages represented by a single speaker.
  - 121 languages with multiple speakers.

Distribution of Speakers by Gender



Top 10 Languages in Speech Accent Archive

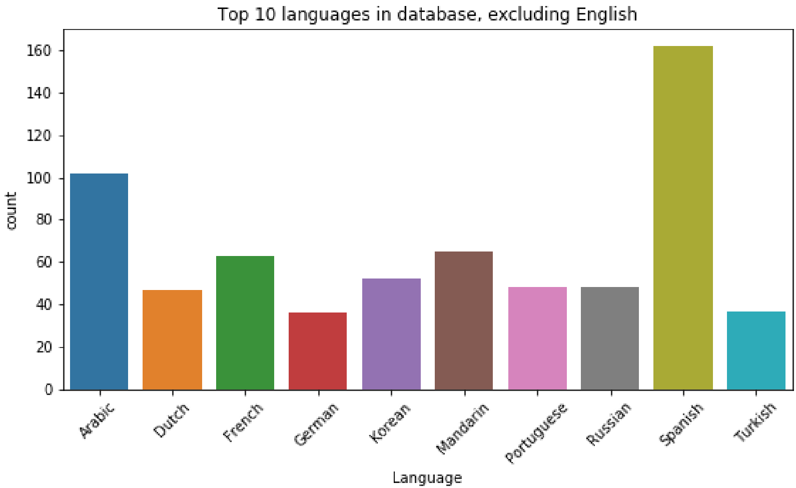


Figure: Number of speakers for top 10 languages, excluding English

Transfer Learning

Transfer learning from image classification models has been an effective starting point for audio classification. [Hershey, et. al., 2017]

VGGish is an audio embedding model based on the architecture of the VGG image classifier, that has been trained on the YouTube-8M Segments Dataset for Acoustic Event Detection Classification (AED). [Hershey, et. al., 2017]

Features extracted by VGGish for AED may be useful for training models for (acoustic) speech classification, with relatively short training times and small datasets.

Model Architecture

Audio files fed through the VGGish model.

- Audio segment (10s) converted to mel spectrogram.
- Mel spectrogram fed through VGGish convolutional layers.
- VGGish output is a 128-D array of features.

VGGish features fed to a Gender Classifier or a Language Classifier.

Classifier models vary in:

- Number of (Dense + Dropout) layer repeats.
- Number of nodes in Dense layer(s).
- Shape of Flatten layer (based on output of previous layer)

Gender Classifier Architecture

Table: Structure of the Gender Classifier models.

Layer	Gender 1	Gender 2
Input	(32, 10, 128)	(32, 10, 128)
Dense Dropout	128 nodes 50%	128 nodes 50 %
Dense Dropout		64 nodes 50%
Flatten	(32, 1280)	(32, 640)
Output	(32, 1)	(32, 1)

Language Classifier Architecture

Table: Structure of the Language Classifier models.

Layer	Language 1	Language 2	Language 3
Input	(32, 10, 128)	(32, 10, 128)	(32, 10, 128)
Dense Dropout	12 nodes 50%	128 nodes 50%	128 nodes 50 %
Dense Dropout			64 nodes 50%
Flatten	(32, 120)	(32, 1280)	(32, 640)
Output	(32, 11)	(32, 11)	(32, 11)

Results

Gender Classifier

Identify the gender of the speaker.

- Convergence after 2-3 epochs.
- 98% accuracy.

Language Classifier

Identify the native language of the speaker.

- 11 language classes.
- Convergence after 7 epochs.
- 24% accuracy with majority class distribution of 16%.

Gender Classifiers - Model Metrics

Table: Summary Metrics

	Gender 1	Gender 2
Loss	0.061640	0.076962
Accuracy	0.981419	0.975507
Precision	0.982818	0.960199
Recall	0.979452	0.991438

Table: Confusion Matrices

		Predicted			Predicted	
Gender 1		F	M	**Gender 2**	F	M
Actual	F	590	10	F	576	24
	M	12	572	M	5	579

Language Classifier - Model Metrics

11 language classes.

Baseline accuracy of 15.7% based on majority class distribution.

		Language 1	Language 2	Language 3
	Loss	2.300652266	2.353183508	2.283754587
	Accuracy	0.178977266	0.23011364	0.241477266
	Precision	1	0.460317463	0.649999976
	Recall	0.002840909	0.082386367	0.082386367
	Micro	0.178977273	0.230113636	0.241477273
F1 score	Macro	0.084855919	0.169606263	0.171586068
	Weighted	0.122333655	0.193686198	0.202591296

Language Classifier - Confusion Matrix

Table: Confusion matrix for Language 3 predictions. The bold numbers on the diagonal represent correct predictions.

Language	R	A	T	K	G	D	S	F	E	P	M	Segments
Russian	**0**	6	3	1	0	3	0	2	1	4	8	28
Arabic	0	**28**	1	0	1	10	1	0	2	0	12	55
Turkish	0	3	**1**	0	0	4	5	1	2	1	4	21
Korean	0	12	0	**2**	0	6	1	1	2	2	5	31
German	0	2	0	1	**0**	1	4	1	4	2	0	15
Dutch	0	1	0	0	0	**17**	1	0	2	2	2	25
Spanish	2	7	0	1	4	3	**6**	4	6	2	7	42
French	0	9	0	0	0	4	4	**3**	0	6	6	32
English	0	2	0	1	0	6	1	5	**9**	0	6	30
Portuguese	0	8	1	0	1	2	1	7	2	**1**	5	28
Mandarin	1	2	3	2	0	5	3	3	2	6	**18**	45
Predictions	3	80	9	8	6	61	27	27	32	26	73	352

Conclusions

- Transfer learning is an effective strategy for training speech models.
- Acoustic features extracted by VGGish for AED are adequate to train a gender classifier with 98% accuracy.
- Language classifier with 11 classes shows improvement over majority class baseline.

Future directions: Optimize language classifier.

- Add more layers and/or nodes.
- Use different activation functions.
- Include additional acoustic features (e.g. tempo/beat tracking).

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., & Zheng, X. 2015. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. https://www.tensorflow.org/.

McFee, B., Lostanlen, V., Metsai, A., McVicar, M., Balke, S., Thomé, C. Raffel, C., Zalkow, F., Malek, A., Dana, Lee, K., Nieto, O., Jack Mason, J., Ellis, D., Battenberg, E., Seyfarth, S. Yamamoto, R. Choi, K., viktorandreevichmorozov, Moore, J. Bittner, R., Hidaka, S., Wei, Z., nullmightybofo, Hereñú, D., Stöter, F., Friesch, P., Weiss, A., Vollrath, M., & Kim, T. 2020. librosa/librosa: 0.8.0. Zenodo, https://doi.org/10.5281/zenodo.3955228

References

Hershey, S. et. al., 2017. CNN Architectures for Large-Scale Audio Classification, *ICASSP*.

Sheng, L. M. A., and Edmund, M. W. X. 2017. Deep Learning Approach to Accent Classification. Project Report Stanford University Stanford CA.

Weinberger, S. 2015. Speech Accent Archive. George Mason University.