MUSIC/SPEECH CLASSIFICATION AND DETECTION MIREX SUBMISSION

Jan Schlüter
Austrian Research Institute for Artificial Intelligence, Vienna
jan.schlueter@ofai.at

ABSTRACT

This submission to the MIREX 2015 Music/Speech Classification and Detection task employs two artificial neural networks to independently detect the presence of music and the presence of speech in a recorded audio signal. They resulted from a 2012 DAFx paper [3], using what were state-of-the-art learning methods at that time. The networks were trained on 15 hours of Swiss radio broadcasts, and achieved accuracies of about 98% on radio recordings. Given their specific focus, they will not work well with signals that are widely different from broadcast radio, but can serve as a baseline to compare against.

1. INTRODUCTION

This document describes both the submission to the detection subtask and the classification subtask of the MIREX 2015 Music/Speech task. The detection method will be described first, followed by an explanation of how the detection results are used to solve the classification task.

2. DETECTION METHOD

As the Music/Speech detector is described in previous work [3], I will just give a brief explanation here. Note that the original work is from 2012, and while the learning methods employed were state-of-the-art at the time, they are unnecessarily complex from a current point of view and obviously miss everything that hadn't been invented yet (such as dropout [1]).

Starting from a mel spectrogram, blocks of 39 frames (covering about 0.9 seconds) are extracted and whitened with PCA, retaining components to cover 99% of the original variance. A mean-covariance Restricted Boltzmann Machine (mcRBM) [2] is trained unsupervisedly to form a generative model of such blocks, and a stack of two regular RBMs is trained on the mcRBM's hidden representation. The weights of the full mcRBM/RBM stack are used to initialize two identical feed-forward neural networks. One of these networks is trained to predict whether the central frame of its input block contains speech, the other is trained to predict music.

At test time, overlapping blocks (with a hop size of 1 frame) are extracted from the mel spectrogram of a recording and processed by the two networks, resulting in a music and speech probability per frame. Both these predictions are smoothed over time with a median filter (of length 250 frames or 5.8 seconds for music, and 100 frames or 2.3 seconds for speech) and thresholded with 0.5. Consecutive stretches of music predictions exceeding the threshold form music segments, and consecutive stretches of speech predictions exceeding the threshold form speech segments. Predictions for music and speech are handled completely independently, so segments can freely overlap.

3. CLASSIFICATION METHOD

Although the classification subtask is devised as a train/test task, my submission does not perform any training whatsoever. Instead, it runs the detector described above and employs a simple heuristic to decide about the class for a given recording: (1) If more than 50% of a recording's frames are classified as speech, classify the recording as speech. (2) Otherwise, if more than 50% are classified as music, classify it as music. (3) Otherwise choose the class more frames were classified as (and speech for a tie).

For what it's worth, this heuristic yields 100% accuracy on the GTZAN music/speech collection.

4. ACKNOWLEDGMENTS

This research is supported by the Austrian Science Fund (FWF): TRP 307-N23.

5. REFERENCES

[1] G.E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R.R. Salakhutdinov. Improving neural networks by preventing co-adaptation of feature detectors. arXiv:1207.0580, 2012.

[2] M. Ranzato and G.E. Hinton. Modeling pixel means and covariances using factorized third-order boltzmann machines. In *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR'10)*, 2010.

[3] Jan Schlüter and Reinhard Sonnleitner. Unsupervised Feature Learning for Speech and Music Detection in Radio Broadcasts. In *Proceedings of the 15th International Conference on Digital Audio Effects (DAFx-12)*, York, UK, 2012.