1. システムの概要

今回の競技部門ではかるたの読みの一部を重ね合わせた音声を扱う。そこで、使用された音声の特徴などを捉えるため、機械学習を用いる。このとき、音声データのままでは機械学習を行うことが困難であるため、問題データに近い音声を作成し、画像に変換することで機械学習を容易にすることを目指す。

以下の図１はシステム概要を簡易的に表したものである。

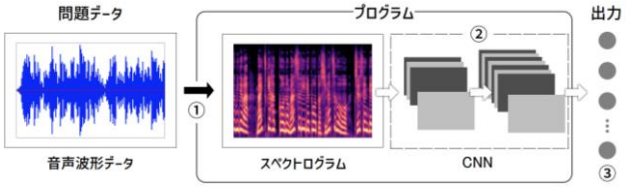

図１．システム概要

2. データセットと学習用画像の作成

機械学習を行うには、入力する画像と正解となるラベルが対応しているデータセットが必要になる。このため、問題データに近い音声を生成するプログラムを作成する。

正解となるラベルは音声とともに生成し、読みデータの総数が日本語と英語がそれぞれ 44 個の計 88 個あるため、1 か 0 の 88 個の整数値の配列とする。

生成した音声からスペクトログラム分析により、周波数と時間を軸にとるグラフを画像として出力する(図１①)

3. 機械学習

畳み込みニューラルネットワーク(CNN)を使用する。これは、入力する画像の特徴を捉えて判別を行うものである。

画像の隣り合ったピクセル同士には波形のように連続性がある。その特徴を生かし、データセットとして作成した各画像に対し、様々なフィルタをかけて画像の枚数を増やした後、ピクセル数を小さくしていくことでどの音声データの特徴を有しているか、識別しやすくする(図１②)。

最終的に機械の予測値は使用された音声が 1,使用されてない数字を 0 に近い値にし、88 個出力する(図１③)。これを正解となるラベルとともに損失関数に渡すことで予測値と正解がどのくらい離れているのかの誤差をとる。誤差が小さくなるほど、正しく予測ができていると判断する。また、１つのデータセットを何度も繰り返し学習させることで、機械が正解となるラベルに近い予測値を出力できるように適応させる。

なお、大会本番で用いる出力は、予測値が指定した値以上であれば 1、指定した値未満であれば 0 の整数値に変換することでどの音声が使用されたのか判断する。