In recent decades, in order to get the utterance level vector representation, dictionary learning procedure is widely used.

A dictionary, which contains several temporal orderless center components (or units, words, clusters), can encode the variable-length input sequence into a single utterance level vector representation.

**LDE Implementation**

The LDE layer is a directed acyclic graph and all the components are parameterized with the input $x = (x_1, x_2, \ldots, x_n)$ and the learnable parameters. Given a set of $L$ frames feature sequence and a learned dictionary center $c = (c_1, c_2, \ldots, c_k)$, each feature of frame $x_i$ can be assigned with a weight to each component and the corresponding residual vector is denoted by $r_i = x_i - \sum_{k=1}^{L} w_{ik} c_k$ where $i = 1, 2, \ldots, n$ and $L = 1, 2, \ldots, m$.

The non-negative assigning weight is given by a softmax function:

$$w_{ik} = \frac{e^{\langle x_i, c_k \rangle}}{\sum_{l=1}^{L} e^{\langle x_i, c_l \rangle}}$$

Given the assignments and the residual vector, conventional GMM Supervisor, the residual encoding model applies an aggregation operation for every dictionary component center $c_k$:

$$x_A = \sum_{i=1}^{n} w_{ik} x_i$$

In order to facilitate the derivation we simplified it as:

$$x_A = \sum_{i=1}^{n} w_{ik} x_i$$

The LDE layer concatenates the aggregated residual vectors with assigned weights. The residual encoder outputs a fixed dimensional representation:

$$E = (x_1, x_2, \ldots, x_n)$$

The task of interest is the closed-set language detection. There are totally 14 target languages in testing corpus, which included 7510 utterances split among three nominal durations: 30, 60 and 3.0 seconds.

In order to get higher abstraction representation better for utterances with long duration, we design a deep CNN based on the well-known ResNet-34 layer architecture, as is described in Table 2. The model parameters of the 3-CNN layer are shown in Table 3. The total parameters of the system performance significantly.

For CNN-TAP system, a simple average pooling layer followed by FC layer is built on top of the front-end CNN. For CNN-LDE system, the average pooling layer is replaced with a LDE layer.

Because we have no separated validation set, even we use the converged model after the last step optimization. For each training step, an interval within (250,000) iterations is randomly initialized, and each data in the mini-batch is cropped or extended to 3 frames. In the testing stage, the 30, 60 and 10 seconds data is tested on the same model. Because the duration length is arbitrary, we feed the testing speech utterance to the trained neural network one by one.

The task of interest is the closed-set language detection. There are totally 14 target languages in testing corpus, which included 7510 utterances split among three nominal durations: 30, 60 and 3.0 seconds.

In order to get higher abstraction representation better for utterances with long duration, we design a deep CNN based on the well-known ResNet-34 layer architecture, as is described in Table 2. The model parameters of the 3-CNN layer are shown in Table 3. The total parameters of the system performance significantly.

For CNN-TAP system, a simple average pooling layer followed by FC layer is built on top of the front-end CNN. For CNN-LDE system, the average pooling layer is replaced with a LDE layer.

Because we have no separated validation set, even we use the converged model after the last step optimization. For each training step, an interval within (250,000) iterations is randomly initialized, and each data in the mini-batch is cropped or extended to 3 frames. In the testing stage, the 30, 60 and 10 seconds data is tested on the same model. Because the duration length is arbitrary, we feed the testing speech utterance to the trained neural network one by one.

The task of interest is the closed-set language detection. There are totally 14 target languages in testing corpus, which included 7510 utterances split among three nominal durations: 30, 60 and 3.0 seconds.

In order to get higher abstraction representation better for utterances with long duration, we design a deep CNN based on the well-known ResNet-34 layer architecture, as is described in Table 2. The model parameters of the 3-CNN layer are shown in Table 3. The total parameters of the system performance significantly.

For CNN-TAP system, a simple average pooling layer followed by FC layer is built on top of the front-end CNN. For CNN-LDE system, the average pooling layer is replaced with a LDE layer.

Because we have no separated validation set, even we use the converged model after the last step optimization. For each training step, an interval within (250,000) iterations is randomly initialized, and each data in the mini-batch is cropped or extended to 3 frames. In the testing stage, the 30, 60 and 10 seconds data is tested on the same model. Because the duration length is arbitrary, we feed the testing speech utterance to the trained neural network one by one.