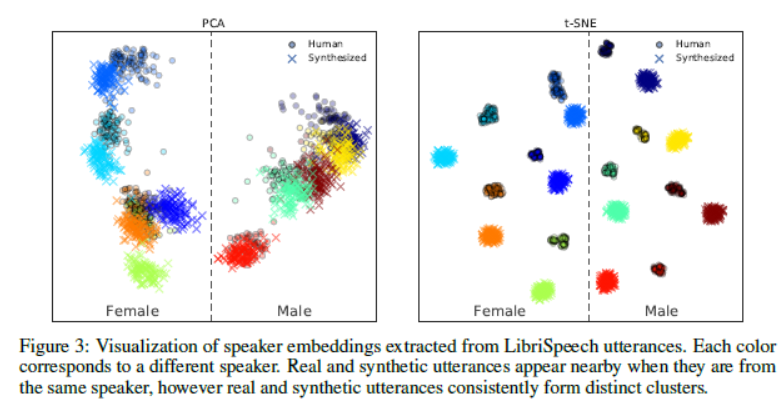
**Goal:** Generate speaker embedding in unsupervised way (no speaker labels in training data)



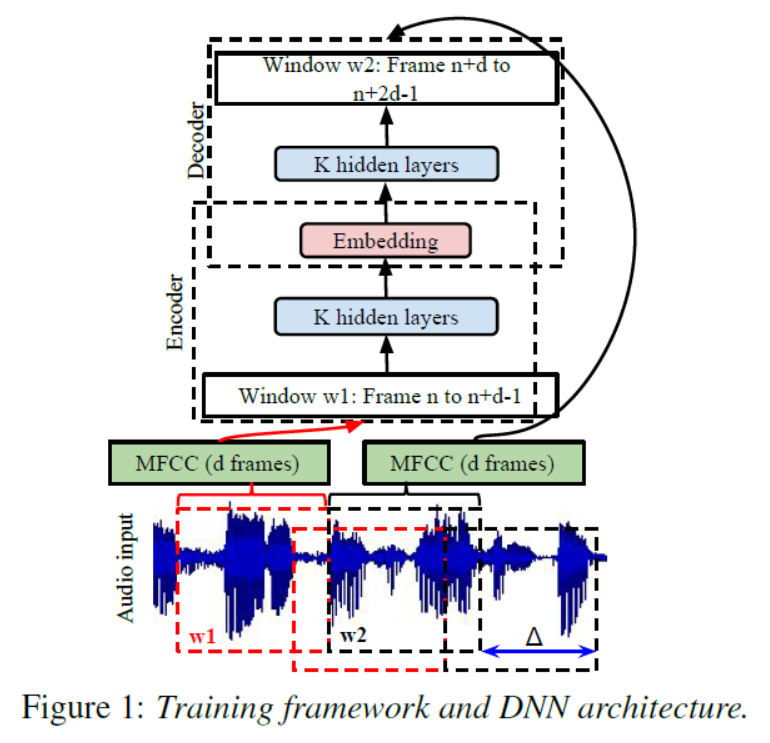
**<Speaker2Vec>**

**Autoencoder architecture**

Input -> K hidden layers -> Embedding layer -> K hidden layers -> Output

Basic: 4000 -> 2000 -> 40 -> 2000 -> 4000

Modified: 4000 -> 6000-> 2000 -> 256 -> 2000 -> 6000 -> 4000



**Input:** A small window of speech (Fixed size, frames , MFCC feature)

**Output:** Fixed size, frames

**Training objective:** Given , reconstruct a small window of speech from a temporally neighboring window (Fixed size, frames , MFCC feature).

**Data:**  a pair. Use all available pairs from an audio input, separated by