TIME-DELAY NEURAL NETWORKS

UCS749: SPEECH PROCESSING AND SYNTHESIS

Raghav B. Venkataramaiyer

CSED TIET Patiala India.

September 16, 2024

Outline

OUTLINE

SUMMARY

In the authors' own words, in the context of "difficult" word recognition, this literature documents the success of,

> *a network that extracted features by repeatedly convolving a set of narrow weight patterns with the contents of a sliding window into the input.*

This is a widely referenced early work that lays foundation for 1-D convolution based approaches in the recent advances, like Jasper and its derivatives.
See also: Dataset to understand the problem better.

OUTLINE

MOTIVATION

> *[...] the Viterbi-aligned speech fragments contained enough alignment errors to motivate a shift-invariant [model]*

Ref: Viterbi Alignment

OUTLINE

KEY IDEA

- Neural networks are universal approximators;
- Convolution controls model size; and
- Pooling leads to positional invariance.

OUTLINE

LET

data	be a set of $(y, \mathbf{x})$ pairs; $\mathbf{x}$ being the input vector, and $y \in \mathbb{Z}_{\geqslant 0}$ being the labels.
$\mathbf{x} \in \mathbb{X} \subseteq \mathbb{R}^{C \times T}$	be input vector; $C = 16$ being number of channels; and T being temporal resolution.
$\Phi(\,\cdot\,;\theta) : \mathbb{X} \to \mathbb{R}^{K \times T'}$	be 1-D convolutional neural network, with its output being a T' long sequence of K dimensional vector, representing word probabilities; K being the vocabulary size and θ being network params.
$\mathbf{1}_k : \mathbb{Z}_{\geqslant 0} \to \mathbb{R}^k$	be whole number to one-hot vector converter
$\lVert\,\cdot\,\rVert_{p,\text{row}}$	be an operator that computes L_p norm for each row vector in the tensor, defined here for mathematical convenience.
$\mathbf{x}^{\otimes n}$	be element-wise exponentiation operation, defined here for mathematical convenience.
Δ_E	be Euclidean distance between two vectors.

MODEL

TDNN model for speech command classification may be summarised mathematically as a learnt function f with optimal parameters θ_*, so that label is predicted as,

$$\widetilde{y} = \arg\max f(\mathbf{x}; \theta_*)$$

Where,

$$\begin{aligned} \theta_* &= \arg\min_{\theta} \mathop{\mathbb{E}}_{y,\mathbf{x}\sim\text{data}} [\Delta(y, f(\mathbf{x}; \theta))] \\ f(\mathbf{x}; \theta) &= \|\Phi(\mathbf{x}; \theta)\|^2_{2,\text{row}} \\ \Delta(y, \widetilde{\mathbf{y}}) &= \Delta_E(\mathbf{1}_K(y), \widetilde{\mathbf{y}}) \end{aligned}$$

OUTLINE

DATASET

400 samples

> *the four words, "bee, dee, ee," and "vee" [B, D, E, and V] were used; earlier IBM research had shown that these four words were the most confusable members of the E-set of the alphabet.*

OUTLINE

OUTLINE

FEATURE EXTRACTION

Initial spectrogram feature

- was extracted from 150 ms waveform
- containing log energy values and
- bearing shape 128×48; 128 frequencies $\times 48$ frames each lasting 3 ms.

In an experiment with input feature size, it was observed, however that

- 16×12 input-based; 16 frequency bands on linear scale $\times 12$ frames each lasting 12 ms
- 2-layer hidden model

exhibited the best performance;

> *[...] the program converted our 150 ms waveform samples into spectrograms containing 128 log energies ranging up to 8 kHz, and 49 time frames of 3 ms each. The first frame of each spectrogram was then discarded so that there would be 48 time steps (a highly factorizable number), and the DC bias com- ponent of each frame was set to zero. Because each of the 48 time frames represented 3 ms, the final duration of the spectrograms was 144 ms.*

(Adapted from the paper)

OUTLINE

VITERBI ALIGNMENT

- In a prior art at IBM, a hidden Markov model (HMM) was used to model the distribution of labels and spoken word.
- The Viterbi search listed the most likely sequence of *labels*, corresponding to each frame of utterance in a spoken word; where the word identity was known.

In such an experiment by the IBM (the authors say)

> *These labels were used to extract a 150 ms salient section of each utterance which included 100 ms before the first frame that was labelled "E" (this region should contain the consonant), plus 50 ms of the vowel.*

HMM MODEL DETAILS (FYI)

Details of the HMM Model in the experiment by IBM

> *[...] the words B, D, and V are modelled by a concatenation of the state machines for noise, voiced consonant onset, {B,D,V}, E, E trail-off, and noise. The word E is modelled by a concatenation of the state machines for noise, E onset, E, E trail-off, and noise. The state machines contain 3 main states with associated transitions to model the beginning, middle, and end of each phone. The consonant and vowel machines include self-loops to model steady-state portions of the acoustic signal, and all of the machines include null transitions to model short durations.*