УДК 004.522

ПРИМЕНЕНИЕ СПЕЦИАЛИЗИРОВАННОЙ НЕЙРОСЕТЕВОЙ АРХИТЕКТУРЫ TDNN ДЛЯ РАСПОЗНАВАНИЯ РЕЧЕВЫХ СИГНАЛОВ

Гусак Е.А., Бондаренко И.Ю.
Донецкий национальный технический университет

Решение проблемы автоматического распознавания речи в настоящее время достаточно актуально по нескольким причинам. Во-первых, систему распознавания и синтеза речи очень удобно использовать для организации еще одного канала управления техникой. В настоящее время используется только тактильно-зрительный канал управления техническими средствами. Соответственно, с расширением функциональных возможностей техники происходит информационная перегрузка этого канала управления, что приводит к переутомляемости оператора и ошибкам в его работе. Создание голосового канала управления позволит распределить информационную нагрузку и избежать подобных проблем. Во-вторых, для людей с ограниченными возможностями, например, с серьезными дефектами зрения, голосовой канал управления является единственной возможностью «общения» с техникой. Реализация системы управления с помощью голосовых команд поможет таким людям эффективно работать с техникой.

Существует множество подходов к реализации системы автоматического распознавания речи. Из них можно выделить нейросетевой подход как один из самых эффективных. Благодаря реализации параллельных вычислений, распознавание речи с помощью нейронной сети происходит значительно быстрее, чем с помощью последовательных алгоритмов. Также нейросеть имеет свойство обучаемости, что позволяет сети самой выделять определенные закономерности и обобщать заданные обучающие образы в классы.

Специализированная нейросетевая архитектура TDNN (Time Delay Neural Network – нейросеть с задержкой времени) была спроектирована специально для распознавания образов, характеризующихся протяженностью и нелинейными изменениями во времени.

Для распознавания речи нейросеть должна обладать такими рядом свойств:

- иметь несколько слоев и достаточное количество взаимосвязей между элементами в каждом слое.
- сеть должна иметь возможность представлять зависимость между событиями во времени. Этими событиями могут быть резонансы на определённых частотах, выраженные спектральными коэффициентами (вместо спектрального анализатора также могут использоваться детекторы признаков более высокого уровня).
- нейросеть должна быть инвариантна к временным сдвигам в распознаваемых речевых образах.
- обучающая процедура не должна быть слишком чувствительной к погрешностям пофонемной сегментации речевых сигналов, на которых происходит обучение.
- количество весов в сети должно быть достаточно малым (пропорционально количеству обучающих данных) для того, чтобы сеть имела способность к обобщению, а не к просто запоминанию обучающих примеров.

TDNN благодаря своей архитектуре обладает всеми вышеперечисленными свойствами.

Архитектура TDNN похожа на архитектуру классического многослойного полносвязного персептрона, но имеет одну важную особенность – в TDNN слои не полностью связаны между собой. Многослойный персептрон – это полносвязная модель без обратных связей. Количество слоев и нейронов в них обычно обосновано постановкой задачи и вычислительными способностями. Количество весовых коэффициентов (синаптических связей) в данной структуре можно вычислить по формуле:

$$l_w = \sum_{i=1}^{N_L - 1} \hat{N}_i \hat{N}_{i-1},$$

где где N_L — количество слоев в ИНС, $\hat{N}_i$ — количество нейронов на i -ом слое. [2]

В TDNN нейрон следующего слоя связан с несколькими нейронами предыдущего слоя, которые называются рецептивным полем нейрона. Это объясняется тем, что нет необходимости обрабатывать одновременно элементы, между которыми существует довольно большой промежуток времени, но есть необходимость обрабатывать одновременно и отслеживать взаимосвязь между довольно близко расположенными друг к другу событиями. Благодаря своей неполносвязности нейросеть с временной задержкой обеспечивает одновременно и эффективную работу обучения и распознавания, и сравнительно малое количество весов.

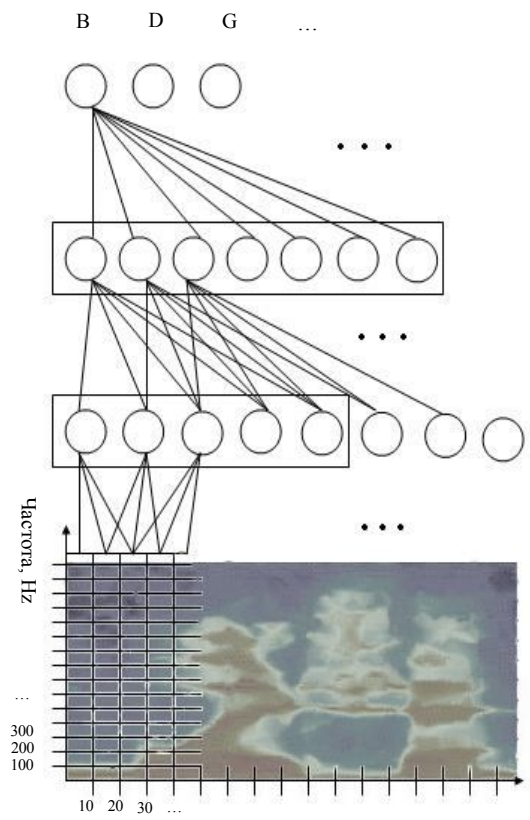

Рис. 1-а. Архитектура TDNN

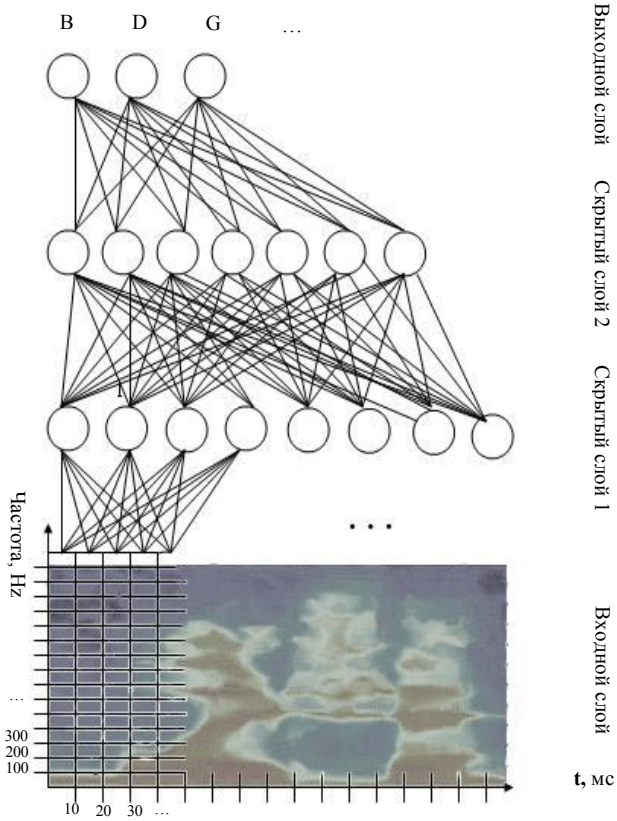

Рис 1-б. Архитектура многослойного полносвязного перспетрона

Основной блок, используемый во многих нейросетях, вычисляет взвешенную сумму входных сигналов и подаёт эту сумму на вход нелинейной функции, обычно пороговой или сигмоиде. В качестве нелинейной функции предпочтительнее сигмоида

из-за своих удобных математических свойств. Входная речь разделена на скользящие окна. Масштабированные коэффициенты вычисляются из спектра мощности посредством вычисления входной мощности в каждой группе масштабированной мощности, где смежные коэффициенты в частоте перекрываются одним спектральным образцом и сглаживаются понижением общего образца на 50 %. Смежные временные коэффициенты сжимаются для дальнейшей редукции данных. Все коэффициенты входящей речи далее нормализуются. Каждый элемент в первом скрытом слое теперь получает входные данные из коэффициентов 3-хфреймового окна. На втором скрытом слое каждый из 3-х элементов TDNN просматривает 5-тифреймовое окно уровней активности в 1-ом скрытом слое. В конце концов, выход получается посредством интеграции (суммирования) данных из каждого из 3-х элементов в скрытом слое 2 в течении времени и присоединения их к подходящему выходному элементу. Каждый элемент TDNN, описанный в этом разделе, имеет возможность кодировать временные отношения внутри диапазона в N задержек. Более высокие уровни могут работать в больших временных промежутках. Процедура обучения обеспечивает то, что каждый элемент в каждом слое имеет вес, отрегулированный таким образом, который улучшает общую производительность сети.

Для обучения этой сети используется метод обратного распространения ошибки. Для достижения желаемой обучаемости должно обеспечиваться получение сетью последовательности данных, чтобы сеть имела возможность запоминать наиболее выраженные сигналы и последовательности сигналов во входном речевом отрезке. Альтернативный вариант достижения этого результата – использовать пространственно расширенный входной образ, т.е., спектрограмму плюс некоторые ограничения в весах. Тогда задача сводится фактически к распознаванию и анализу изображения спектрограммы.

Исходя из результатов исследований, проведенных авторами статьи [1], можно сделать вывод о том, что нейросеть типа TDNN является наиболее перспективной нейросетевой архитектурой для решения задачи распознавания речи, так как показывает серьезные результаты распознавания и высокий процент верных результатов при сравнительно меньших требованиях к вычислительной мощности, в отличие от альтернативных методов. В нейросети типа TDNN можно реализовать большее количество слоев, что обеспечивает инвариантность к различным искажениям, и при этом количество весов за счет неполносвязности остается небольшим, что упрощает обучение. Дальнейшие исследования будут направлены на разработку программно-аппаратной модели нейросети с задержкой времени и эксперименты по её применению в задаче распознавания речевых сигналов.

Литература

[1] A. Waibel, T. Hanazawa, G. Hinton, K.Shikano, K.J. Lang. Phoneme Recognition Using Time-Delay Neural Networks – IEEE Transaction on acoustic, speech, and signal processing. Vol. 37 No. 3 March 1989 – p. 328

[2] Крючин О.В., Козадаев А.С., Дудаков В.П. Прогнозирование временных рядов с помощью искусственных нейронных сетей и регрессионных моделей на примере прогнозирования котировок валютных пар. - Электронный научный журнал «ИССЛЕДОВАНО В РОССИИ» - с. 354