УДК 004.85

Т.А. Брынза, Н.Е. Губенко
Донецкий национальный технический университет, г. Донецк
кафедра компьютерных систем мониторинга

ИСПОЛЬЗОВАНИЕ БАЙЕСОВСКИХ СЕТЕЙ ДОВЕРИЯ В ПРИМЕНЕНИИ К РАСПОЗНАВАНИЮ РЕЧИ

Аннотация

Брынза Т.А, Губенко Н.Е., Использование байесовских сетей доверия в применении к распознаванию речи. *Рассмотрены подходы к представлению байесовских сетей доверия для распознавания речи. Указаны особенности данной задачи и способ ее решения с помощью байесовских сетей доверия. Приведены варианты использования данной нейросетевой архитектуры в акустико-фонетической системе автоматического распознавания речи.*

Ключевые слова: *распознавание речи, байесовские сети доверия, автоматические системы распознавания речи, ограниченная машина Больцмана, акустическая модель*

Постановка проблемы. Машинное распознавание речи является важным объектом исследований уже на протяжении более десяти лет. Прогресс, достигнутый за последние годы, впечатляет, но все еще автоматизированные системы распознавания речи не могут на равных конкурировать со способность распознавать речь человеческим мозгом. Поэтому поиск такой архитектуры, способной решить задачу распознавания речи, которая бы была способна справиться с этой задачей наилучшим образом, представляется актуальной задачей.

Анализ литературы. Поиску оптимальной архитектуры для распознавания речи посвящено большое количество научных публикаций. И очень часто в качестве акустического классификатора авторы выбирают именно байесовские сети доверия. Среди большого количества таких исследований следует выделить статьи Хинтона и соавторов [1], [2], Мохри и соавторов [3], и Мантавона [4]. В [1] Хинтон и его коллеги и Мантавон в [4] показали, что в целом многослойные нейронные сети справляются с задачей акустического моделирования лучше, чем, например, гауссова модель смешения, и что данная архитектура устойчива к шуму и обладает лучшим быстродействием. Абдель-рахман, Хинтон и Пенн в [2] предложили для распознавания речи использовать байесовские сети доверия (DBN), одну из архитектур многослойных нейронных сетей. Их исследование показало, что данная архитектура является хорошим классификатором благодаря тому, что

большое количество скрытых слоев позволяет выявить нелинейные признаки, а предобучение в качестве генеративной модели позволяет упростить и ускорить процесс обучения сети. Также в этой статье предлагается рассматривать скрытые слои DBN как стек ограниченных машин Больцмана (RBM), что позволяет получить достоверную выборку из апостериорного распределения по данному вектору входных данных, а это в свою очередь упрощает обучение сети. В 2008 Мохри и соавторы показали, что для наиболее эффективного распознавание речи лучше всего использовать гибридные акустико-фонетические модели, так как это позволяет использовать контекст предложения для более точного распознавания сказанного. В предложенной авторами гибридной модели DBN выполняют роль акустического классификатора, а скрытые Марковские модели (HMM) отвечают за фонетическое распознавание. Данный подход набирает все большую популярность, и работа [3] задает тенденцию в распознавании речи.

Цели статьи – изучение байесовских сетей доверия в контексте распознавания речи и выявление такой конфигурации данной модели, которая бы решала задачу акустического моделирования наилучшим образом.

Представление байесовских сетей доверия для распознавания речи

Как было указанно в [5], существует несколько подходов к представлению байесовских сетей доверия. В данной статье мы остановимся только на одном из них – многослойной нейросетевой архитектуре, так как анализ литературы показал, что именно такое представление является наиболее удачным в контексте задачи распознавания речи.

Если абстрагироваться от деталей, то сети доверия очень сильно схожи с многослойными персептронами. Вычисления условной вероятности активации переменной в ответ на собственное индуцированное локальное поле в байесовских сетях определяется следующей функцией [6]:

$$P\left(X_j = x_j \middle| pa\left(X_j\right)\right) = \varphi\left(x_j \sum_{i<j} w_{ji} x_i\right) \qquad (1)$$

где x_j – это переменная, определяющая сеть доверия, а w_j – вероятностная зависимость между двумя переменными.

Недостаток байесовских сетей доверия состоит в том, что крайне сложно получить апостериорное распределение по всем возможным конфигурациям скрытых элементов сети, а в случае с распознавание речи, где параметров очень много, это практически невозможно. Обучение было бы простым, если бы удалось получить достоверную выборку по апостериорному распределению скрытых состояний по наблюдаемым данным [2]. Эта процедура является достаточно сложной в том числе и в следствии «эффекта объяснения», который заключается в том, что две причины (элементы скрытого слоя сети доверия), являющиеся независимыми, могут быть зависимыми для того, кто наблюдает некоторый эффект, который зависит от

обеих этих причин. Таким образом, чтобы получить апостериорное распределение даже в первом скрытом слое сети, требуется интегрировать по всем возможным конфигурациям скрытых переменных. А как было описано выше, эта процедура в сложных сетях становится невозможной.

Обойти такой недостаток DBN можно, если, следуя рекомендациям, изложенным в [2], в качестве скрытого слоя сети использовать стек RBM.

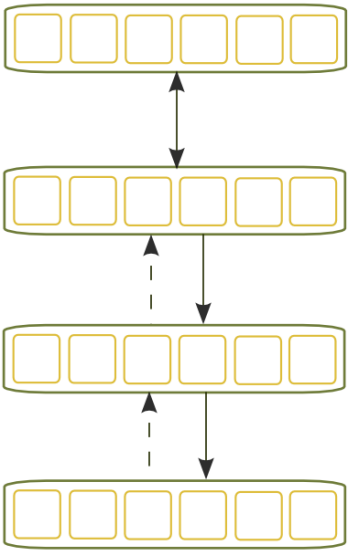

Рисунок 1 – Модель ограниченной машины Больцмана

А ограничив некоторым образом соединения, можно добиться упрощения в обучении данной архитектуры [7]. Это ограничение состоит в том, чтобы обучать сеть по одному слою скрытых элементов за раз. Позднее сеть будет доучена уже со всеми слоями сразу, но для предобучения такая методика является очень эффективной, так как все элементы в одном скрытом слое условно независимы для данных видимых состояний.

Особенности решения задачи распознавания речи с помощью байесовских сетей доверия

Так как мы рассматриваем DBN в контексте распознавания речи, то стоит отметить несколько важных аспектов решения данной задачи.

В первую очередь следует отметить, феномен множества акцентов, что затрудняет автоматическое распознавание речи. Одним из возможных вариантов решения данной проблемы представляется обучение сети с помощью записи голоса дикторов из разных стран и регионов. Например, в речевом корпусе TIMIT, который общепризнанным корпусом для обучения

систем распознавания речи, используется 630 американских дикторов, которые разделены на 8 «региональных диалектов». Это позволяет проводить процедуру обучения более эффективно.

Также важным аспектом является тот факт, что в автоматическом распознавании речь представляется цифровым сигналом акустической волны, в которой «речевые события» плавно перетекают друг в друга, не образуя четких границ [8]. И поэтому, в автоматическом распознавании речи оперируют понятием «фрейм» - небольшим участком исходного сигнала речи. Такой подход позволяет выделить отдельные фонемы и распознать их по отдельности. Но проблема заключается в том, что многие фонемы чрезвычайно похожи друг на друга. Но если нельзя дать однозначный ответ, то можно рассуждать в терминах «вероятностей»: для данного сигнала одни фонемы более вероятны, другие менее, третьи вообще можно исключить из рассмотрения. Собственно, акустическая модель — это функция, принимающая на вход небольшой участок акустического сигнала (фрейм) и выдающая распределение вероятностей различных фонем на этом фрейме. Таким образом, акустическая модель дает нам возможность по звуку восстановить, что было произнесено — с той или иной степенью уверенности. Для построения такой модели рекомендуется использовать байесовские сети доверия, так как рассуждение в них ведется в терминах теории вероятности (1).

Последней особенностью задачи распознавания речи, на которой заострим внимание, является использование контекста для улучшения распознавательных качеств автоматизированной системы. Действительно, люди лучше машин справляются с распознаванием речи так как умеют оперировать контекстом фразы, и всего сообщения в целом. Поэтому даже не расслышав некоторое слово, мы можем восстановить ее своих знаний п представлений о теме беседы. Компьютерные модели распознавания речи лишены такой возможности. Для компенсации данного недостатка рекомендуется решать задачу распознавания речи с помощью акустико-фонетических систем. В [4, 9] в качестве фонетического компонента распознавания устной речи предлагают использовать HMM. Данный подход нашел практическое применение, и используется, например, в системе распознавания речи компании Yandex. Такое решение имеет ряд достоинств и недостатков [9]. Мы же скажем только, что имеется и альтернативное решение методами динамического программирования, основанных детерминированных, а не, как в случае с HMM, вероятностных вычислениях. Выбор архитектуры для фонетического распознавания требует дальнейших исследований.

Выводы. Байесовские сети доверия являются одной из найболее удачных архитектур распознавания речи, так как такая архитектура в своей основе содержит матапарат теории вероятностей, то очень хорошо справляется с

распознаванием устной речи. В применении к этой задаче, DBN рекомендуется рассматривать как стек ограниченных машин Больцмана.

Для максимально эффективного распознавания устной речи следует придерживаться таких рекомендаций: чтобы сделать систему независимой от акцентов, следует обучать сеть на речевом корпусе, который содержит речь дикторов из разных регионов, например, TIMIT. Также стоит разбивать речь на небольшие фреймы и распознавать отдельные фонемы, а за тем переходить на более обобщенный уровень распознавания. Для того, чтобы иметь возможность распознать речь из контекста, следует в дополнении к DBN использовать или HMM, или методы динамического программирования.

Литература

1. Deng, L., Hinton, G. E. and Kingsbury, B. New types of deep neural network learning for speech recognition and related applications: An overview – IEEE International Conference on Acoustic Speech and Signal Processing (ICASSP 2013) – Vancouver, 2013. – 5 p.

2. Abdel-rahman Mohamed, Geoffrey Hinton, Gerald Penn. Understanding how Deep Belief Nets perform acoustic modeling. – ICASSP, 2012 – 4 p.

3. Mohri, M., Pereira, F., & Riley, M. Speech recognition with weighted finite-state transducers. – In Springer Handbook of Speech Processing. Springer Berlin Heidelberg, 2008 – 25 p.

4. Gregoire Montavon. Deep learning for spoken language identification. – Machine Learning Group, Berlin Institute of Technology Germany, 2005 – 4 p.

5. Брынза Т.А., Бондаренко И.Ю., Губенко Н.Е. Представление байесовских сетей доверия для решения задачи распознавания образов. – Труды IX международной научно-технической конференции студентов, аспирантов, молодых ученых «Информатика и компьютерные технологии», 2013 – 4 ст.

6. Брынза Т.А, Бондаренко И.Ю. Сигмоидальные сети доверия в решении задач классификации – Труды IV международной конференции «Информационно-управляющие системы и компьютерный мониторинг», 2013 – 6 ст.

7. Linda Otmani, Abdelkader Benyettou. Les réseaux neuro-bayésiens appliqués à la reconnaissance de la parole. – Université des sciences et de technologie d'ORAN -Mohamed Boudiaf- faculté des sciences, département d'informatique, 2007 – 7 p.

8. Распознавание речи от Яндекса. Под капотом у Yandex.SpeechKit. Электронный ресурс. Режим доступа: http://habrahabr.ru/company/yandex/blog/198556/

9. Hinton, G., Deng, L., Yu, D., Dahl, G. E. et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. – Signal Processing Magazine, IEEE, 2012 – 11 p.