(19) Евразийское патентное ведомство

(11) 042412 (13) B1

(12) ОПИСАНИЕ ИЗОБРЕТЕНИЯ К ЕВРАЗИЙСКОМУ ПАТЕНТУ

(45) Дата публикации и выдачи патента
2023.02.10

(21) Номер заявки
202092400

(22) Дата подачи заявки
2018.05.03

(51) Int. Cl. ***G10L 15/02*** (2006.01)
G10L 17/18 (2013.01)
G10L 15/07 (2013.01)
G10L 15/16 (2006.01)

(54) СПОСОБ ПОЛУЧЕНИЯ ДИКТОРОЗАВИСИМЫХ МАЛОРАЗМЕРНЫХ ВЫСОКОУРОВНЕВЫХ АКУСТИЧЕСКИХ ПРИЗНАКОВ РЕЧИ

(43) **2021.03.03**
(86) **PCT/RU2018/000286**
(87) **WO 2019/212375 2019.11.07**

(71)(73) Заявитель и патентовладелец:
ОБЩЕСТВО С ОГРАНИЧЕННОЙ ОТВЕТСТВЕННОСТЬЮ "ЦЕНТР РЕЧЕВЫХ ТЕХНОЛОГИЙ" (RU)

(72) Изобретатель:
Прудников Алексей Александрович, Кореневский Максим Львович, Меденников Иван Павлович (RU)

(74) Представитель:
Нилова М.И. (RU)

(56) WO-A1-2017099936
US-B2-9858919
MEDENNIKOV I.P. Metody, algoritmy i programmnye sredstva raspoznavaniya russkoi telefonnoi spontannoi rechi. Dissertatsiya na soiskanie uchenoi stepeni kandidata tekhnicheskikh nauk. Sankt-Peterburg, 2016

(57) Изобретение относится к области распознавания речи, в частности к получению высокоуровневых акустических признаков речи для распознавания речи в условиях акустической вариативности. Предложен способ получения малоразмерных высокоуровневых акустических признаков речи, согласно которому обеспечивают наличие низкоуровневых признаков речи и соответствующей им дикторской информации, затем обучают нейронную сеть с использованием низкоуровневых признаков речи, после чего дообучают нейронную сеть с использованием низкоуровневых признаков речи, дополненных дикторской информацией. Вводят малоразмерный слой в состав нейронной сети и дообучают нейронную сеть с малоразмерным слоем с использованием низкоуровневых признаков речи, дополненных дикторской информацией, затем извлекают с выхода малоразмерного слоя нейронной сети малоразмерные высокоуровневые акустические признаки речи.

042412 B1

042412 B1

Область техники

Изобретение относится к области распознавания речи, в частности к получению высокоуровневых акустических признаков речи для распознавания речи в условиях акустической вариативности.

Уровень техники

Известен способ получения дикторoадаптивной акустической модели посредством нейронной сети с использованием i-вектора (US2015149165). Согласно известному способу, обеспечивают наличие акустической модели на основе глубокой нейронной сети, принимают аудиоданные, включающие одно или несколько высказываний диктора, извлекают множество признаков распознавания речи из указанных одного или нескольких высказываний диктора, создают идентификационный вектор диктора для этого диктора на основе извлеченных признаков распознавания речи и адаптируют акустическую модель глубокой нейронной сети для автоматического распознавания речи с использованием извлеченных признаков распознавания речи и идентификационного вектора диктора.

Известен способ адаптации акустической модели на основе нейронной сети (US20170169815). В одном из вариантов реализации способа обученная акустическая нейронная сеть может быть адаптирована к диктору путем использования речевых данных, соответствующих множеству высказываний, произносимых диктором. Обученная нейронная сеть акустической модели может иметь входной слой, один или несколько скрытых слоев и выходной слой и может быть глубокой нейронной сетью. Входной слой может включать в себя набор входных узлов, которые содержат признаки речи, полученные из речевого высказывания, и другой набор входных узлов, которые содержат значения информации о дикторах, полученные из речевого высказывания. Признаки речи могут включать значения, используемые для сбора информации о содержании произносимого высказывания, включая, без ограничения, мел-частотные кепстральные коэффициенты (MFCCs) для одного или нескольких речевых кадров, производные первого порядка между MFCCs для последовательных речевых кадров (delta MFCCs) и производные второго порядка между MFCCs для последовательных кадров (delta-delta MFCCs). Кроме того, значения информации о дикторе могут включать вектор идентификации диктора (i-вектор).

Общим недостатком известных способов (US2015149165 и US20170169815) является то, что они не обеспечивает получения акустической модели, которая бы обладала высокой устойчивостью к искажениям входных данных и позволяла бы с высокой точностью распознавать речь в условиях акустической вариативности.

Известна многоязычная акустическая нейронная сеть (US9460711). В данном документе описывается система многозадачного обучения. Акустическая модель с многоязычной глубокой нейронной сетью может иметь нейронную сеть прямого распространения, имеющую несколько слоев с одним или несколькими узлами. Каждый узел данного слоя соединен соответствующими весами с каждым узлом последующего слоя, а несколько слоев с одним или несколькими узлами могут иметь один или несколько общих скрытых слоев узлов и языкозависимый выходной слой узлов, соответствующих каждому из двух или более языков.

Недостатками известного изобретения является то, что раскрытый в нем способ не обеспечивает получение многоязычной акустической модели, которая бы обладала высокой устойчивостью к искажениям входных данных и позволяла бы с высокой точностью распознавать речь в условиях акустической вариативности.

Известен способ распознавания речи с использованием нейронной сети с адаптацией к диктору (US9721561), согласно которому проводят обучение акустической модели на основе глубокой нейронной сети с узким горлом (с малоразмерным слоем), на вход которой поступают акустические признаки и дополнительная дикторская информация, благодаря чему осуществляется дикторосведомленное обучение. Согласно одному из вариантов осуществления способа, глубокая нейронная сеть имеет несколько скрытых слоев, малоразмерный слой и выходной слой, при этом ее первый скрытый слой включает в себя первый набор узлов, обрабатывающий акустические признаки, и второй набор узлов, обрабатывающий дополнительную дикторскую информацию, входные акустические признаки умножаются на первую матрицу весовых коэффициентов, а дополнительная дикторская информация умножается на вторую матрицу весовых коэффициентов. Выходы малоразмерного слоя соединены со следующим слоем сети.

Недостатком известного способа является то, что обучение нейронной сети с узким горлом не обеспечивает получения качественных малоразмерных признаков и, как следствие, не может обеспечить получения акустической модели, которая бы позволяла с высокой точностью распознавать речь в условиях акустической вариативности.

Известен способ пополнения данных, основанный на стохастическом преобразовании признаков для автоматического распознавания речи (US9721559), согласно которому обучают дикторозависимую акустическую модель целевого диктора для дальнейшего распознавания его речи, т.е. указанная модель призвана с наилучшим возможным качеством распознавать одного конкретного диктора. Ввиду недостаточности данных целевого диктора для обучения нейронной сети предлагается дополнять имеющиеся данные данными других дикторов из обучающей выборки, преобразованными с помощью стохастического преобразования признаков, а также возмущения длины голосового тракта. Параметры данных преобразований оцениваются на основе первой акустической модели, построенной только по данным целе-

вого диктора. После дополнения выборки производится двухэтапное обучение: на первом этапе обучают глубокую нейронную сеть с узким горлом (с малоразмерным слоем) для получения признаков, которые извлекаются из малоразмерного слоя и используются во втором этапе обучения нейронной сети для получения результирующей дикторозависимой модели.

Недостатком известного способа является то, что малоразмерные признаки строятся для конкретного диктора и используются для обучения акустической модели, предназначенной исключительно для распознавания его речи. Полученные таким способом признаки не позволяют обучить акустическую модель, которая бы использовалась для распознавания речи произвольных дикторов. Кроме того, предложенный способ обучения нейронной сети с узким горлом не обеспечивает получения качественных малоразмерных признаков.

Таким образом, известные способы не обеспечивают получения акустических признаков и акустических моделей, отвечающих высокому уровню качества, для последующего распознавания речи в условиях акустической вариативности различных дикторов. Кроме того, недостаточно проработана возможность получения многоязычных акустических признаков и/или многоязычной акустической модели, обладающих высокой устойчивостью к искажениям входных данных и отвечающих высокому уровню качества для последующего распознавания речи.

Ввиду имеющихся недостатков известных способов получения акустических признаков и/или акустических моделей технической проблемой настоящего изобретения является создание способа получения высокоуровневых акустических признаков, которые могут быть использованы для обучения акустической модели, характеризующейся низкой чувствительностью к акустической вариативности речевого сигнала и обеспечивающей высокую точность при распознавании речи.

Раскрытие сущности изобретения

Поставленная проблема решена благодаря тому, что, согласно предлагаемому способу получения малоразмерных высокоуровневых акустических признаков речи, обеспечивают наличие низкоуровневых признаков речи и соответствующей им дикторской информации, затем обучают нейронную сеть с использованием низкоуровневых признаков речи, после чего дообучают нейронную сеть с использованием низкоуровневых признаков речи, дополненных дикторской информацией. Далее, вводят малоразмерный слой в состав нейронной сети и дообучают нейронную сеть с малоразмерным слоем с использованием низкоуровневых признаков речи, дополненных дикторской информацией, затем извлекают с выхода малоразмерного слоя нейронной сети малоразмерные высокоуровневые акустические признаки речи.

Предлагаемый способ позволяет достичь технического результата в виде повышения информативности высокоуровневых акустических признаков, что, в свою очередь, позволяет повысить точность систем распознавания речи различных (произвольных) дикторов в условиях акустической вариативности.

Согласно предлагаемому способу, помимо низкоуровневых речевых признаков используют дополнительную дикторскую информацию (например, с использованием i-векторов), учитывающую информацию о дикторе, и/или канале, и/или окружении, которая позволяет получить так называемые дикторозависимые акустические признаки, обеспечивающие распознавание речи различных (произвольных) дикторов и в различных условиях. Реализация предлагаемого способа осуществляется на основе нейронных сетей, что позволяет повысить качество получаемых акустических признаков. В предлагаемом способе используют нейронную сеть с узким горлом, т.е. вводят в нейронную сеть малоразмерный слой, что понижает размерность входных данных. Кроме того, после обучения нейронной сети выходы этого слоя будут представлять собой малоразмерные высокоуровневые признаки, не только устойчивые к искажениям входных акустических признаков, но также аккумулирующие в себе информацию о дикторе, и/или канале, и/или окружении. Стоит отметить, что качество обучения нейронной сети напрямую влияет на качество получаемых в результате обучения дикторозависимых малоразмерных высокоуровневых акустических признаков.

В предлагаемом способе первоначальное обучение нейронной сети производят с использованием только низкоуровневых речевых признаков, а затем с использованием низкоуровневых речевых признаков, дополненных дикторской информацией, без малоразмерного слоя, что позволяет привести веса остальных слоев к значениям, достаточно близким к оптимальным, что повышает качество обучения нейронной сети и облегчает дообучение сети после внедрения малоразмерного слоя. Дообучение нейронной сети с использованием низкоуровневых речевых признаков, дополненных дикторской информацией, позволяет компенсировать изменения в матрице весов последнего слоя после ввода в нейронную сеть малоразмерного слоя, что повышает качество обучения нейронной сети и, как следствие, качество получаемых после обучения акустических признаков. Использование дикторозависимых малоразмерных высокоуровневых акустических признаков, полученных предложенным способом, для обучения нейронной сети распознаванию речи позволяет получить существенные приросты в точности распознавания речи.

Согласно частному случаю реализации, после обучения нейронной сети с использованием низкоуровневых речевых признаков ее входной слой расширяют путем дополнения матрицы слоя нулевыми столбцами. Расширение входного слоя необходимо для обеспечения возможности дообучения нейронной сети с использованием низкоуровневых речевых признаков, дополненных дикторской информацией, в противном случае размерность входного вектора, состоящего из низкоуровневых речевых признаков и соответствующей им дикторской информации, будет слишком велика для входного слоя нейронной сети.

Кроме того, расширение путем дополнения матрицы входного слоя нулевыми столбцами после обучения нейронной сети с использованием низкоуровневых речевых признаков позволяет сохранить поведение сети, что улучшает качество обучения нейронной сети.

Согласно частному случаю реализации, низкоуровневые речевые признаки имеют вид мел-частотных кепстральных коэффициентов либо логарифмов энергии в мел-частотных полосах. Представление низкоуровневых речевых признаков в предложенных видах позволяет обеспечить получение качественных высокоуровневых акустических признаков.

Согласно частному случаю реализации, дикторская информация имеет вид малоразмерного i-вектора. I-вектор представляет собой малоразмерный (порядка 100 элементов) вектор, который позволяет кодировать отклонение распределения акустических признаков фонограммы от распределения, оцененного по всей обучающей выборке, и аккумулировать в себе информацию о дикторе, а также, в некоторой степени, о канале и акустическом окружении. Таким образом, использование малоразмерного i-вектора совместно с низкоуровневыми речевыми признаками повышает точность обучения нейронной сети и, как следствие, получаемых в результате обучения высокоуровневых акустических признаков.

Согласно частному случаю реализации, обучение нейронной сети с использованием низкоуровневых речевых признаков проводят по критерию минимума кросс-энтропии. Кросс-энтропия показывает, насколько распределение вероятностей на выходе нейронной сети соответствует реально наблюдаемому на данном кадре сенону. Таким образом, использование данного критерия повышает точность обучения нейронной сети.

Согласно частному случаю реализации, дообучают нейронную сеть с использованием низкоуровневых речевых признаков, дополненных дикторской информацией, по критерию минимума суммы кросс-энтропии и дополнительного регуляризирующего слагаемого. Дополнительное регуляризирующее слагаемое препятствует сильному отклонению весов от ранее обученных, что увеличивает качество (точность) обучения нейронной сети.

Согласно частному случаю реализации, нейронную сеть, обученную с использованием низкоуровневых речевых признаков, дополненных дикторской информацией, по критерию минимума суммы кросс-энтропии и дополнительного регуляризирующего слагаемого дообучают по последовательно-дискриминативному критерию. Данный критерий повышает точность распознавания.

Согласно частному случаю реализации, вводят малоразмерный слой путем низкоранговой факторизации матрицы весов последнего скрытого слоя, в частности путем сингулярного разложения. Сингулярное разложение позволяет снизить ранг матрицы весов последнего скрытого слоя нейронной сети путем отбрасывания наименьших сингулярных чисел, тем самым обеспечивая ввод в нейронную сеть малоразмерного слоя (малоразмерного линейного слоя).

Согласно частному случаю реализации, после завершения дообучения нейронной сети с малоразмерным слоем слои, находящиеся после малоразмерного слоя нейронной сети, удаляют. Удаление всех слоев после малоразмерного слоя позволит рассматривать обученную нейронную сеть как экстрактор малоразмерных высокоуровневых признаков.

Согласно частному случаю реализации, подают низкоуровневые речевые признаки по меньшей мере двух различных языков и соответствующую им дикторскую информацию на вход нейронной сети и извлекают с выхода малоразмерного слоя нейронной сети многоязычные малоразмерные высокоуровневые акустические признаки речи. После обучения нейронной сети предложенным выше способом с использованием различных языков из обучающей выборки малоразмерный слой содержит в себе высокоуровневые признаки, относящиеся ко всем языкам обучающей выборки сразу. Полученные таким образом акустические признаки имеют высокую информативность и могут повысить устойчивость к изменению языка входных данных в системах распознавания речи.

Согласно частному случаю реализации, количество выходных слоев нейронной сети равно количеству языков, при этом веса каждого из выходных слоев настраивают только по данным соответствующего языка, а веса всех скрытых слоев настраивают по данным всех из указанных по меньшей мере двух языков. Предложенная архитектура обеспечивает возможность многоязычного обучения нейронной сети.

Краткое описание чертежей

Сущность изобретения более подробно поясняется на неограничительных примерах его осуществления со ссылкой на прилагаемые чертежи, среди которых:

фиг. 1 - архитектура обучаемой нейронной сети без малоразмерного слоя, согласно одному из вариантов осуществления изобретения;

фиг. 2 - архитектура обучаемой нейронной сети с малоразмерным слоем, согласно одному из вариантов осуществления изобретения;

фиг. 3 - схема обучения нейронной сети распознавания речи, согласно одному из вариантов осуществления изобретения.

Осуществление изобретения

Одной из наиболее сложных задач в области автоматического распознавания речи является проблема распознавания разговорной спонтанной речи различных (произвольных) дикторов. Сложность задачи обусловлена особенностями разговорной спонтанной речи различных (произвольных) дикторов: высокие канальная и дикторская вариативность, наличие аддитивных и нелинейных искажений, наличие акцент-

ной и эмоциональной речи, разнообразная манера произнесения, вариативность темпа речи, редукция и вялая артикуляция. Одним из способов повышения качества распознавания спонтанной речи является снижение чувствительности системы распознавания к акустической вариативности речевого сигнала. Реализация данного способа возможна при применении адаптации акустических моделей на основе глубоких нейронных сетей с использованием дикторской информации, учитывающей информацию о диктore, и/или канале, и/или окружении.

Предложенный согласно различным вариантам реализации способ получения малоразмерных высокоуровневых акустических признаков речи позволяет получить акустические признаки, которые могут быть использованы для обучения адаптивной акустической модели, характеризующейся низкой чувствительностью к акустической вариативности речевого сигнала и обеспечивающей высокую точность при распознавании речи.

Подробная последовательность операций способа получения малоразмерных дикторозависимых высокоуровневых признаков речи, согласно одному из вариантов реализации изобретения, раскрыта ниже.

В настоящем описании под термином "дообучение" понимается обучение, начинающееся с настроенных параметров, полученных в ходе предыдущего обучения.

Способ получения малоразмерных высокоуровневых акустических признаков речи в соответствии с настоящим изобретением может быть осуществлен с использованием, например, известных компьютерных или мультипроцессорных систем. В других вариантах реализации заявленный способ может быть реализован посредством специализированных программно-аппаратных средств.

Для получения дикторозависимых высокоуровневых признаков используют глубокую нейронную сеть прямого распространения. В других вариантах реализации могут быть использованы другие подходящие архитектуры для обучения нейронной сети, например сверточные нейронные сети, нейронные сети с задержкой по времени и т.д. Базовую глубокую нейронную сеть прямого распространения изначально инициализируют случайными весами, после чего подают на ее вход обучающий пример и вычисляют активность сети, затем формируют представление об ошибке, то есть разность между тем, что должно быть на выходном слое, и что получилось у сети. Далее веса корректируют таким образом, чтобы уменьшить эту ошибку.

На фиг. 1 изображена глубокая нейронная сеть прямого распространения без малоразмерного слоя (без узкого горла). Предложенная нейронная сеть содержит входной слой 1, на который подают низкоуровневые признаки речи и i-вектор. Также нейронная сеть содержит несколько скрытых слоев 2, которые обрабатывают признаки, полученные с входного слоя, и выходной слой 3, который выводит результат. Каждый слой содержит нейроны, которые получают информацию, производят вычисления и передают ее дальше. Между нейронами есть связи - синапсы, которые имеют параметр - вес, благодаря которому входная информация изменяется в процессе передачи от одного нейрона к другому, при этом совокупность весов нейронной сети образуют матрицу весов. В процессе обучения нейроны изменяют весовые коэффициенты; иными словами, весовые коэффициенты нейронов изменяются с учетом информации, поступающей на нейрон. Изначально обучение проводят посредством глубокой нейронной сети без узкого горла (без малоразмерного слоя), после обучения нейронной сети до необходимых пределов добавляют малоразмерный слой 2a (фиг. 2).

Для получения дикторозависимых высокоуровневых признаков используют глубокую нейронную сеть прямого распространения, обучаемую для классификации единиц речи. На каждом кратковременном участке речи (кадре, обычно они следуют с частотой 100 Гц) классификация позволяет оценить, какими произнесенными "звуками" речи вероятнее всего был порожден наблюдаемый вектор акустических признаков. Под единицами речи могут пониматься фонемы. В настоящем описании термин "фонема" означает минимальную единицу звукового строя языка, не имеющую самостоятельного лексического или грамматического значения. Например, согласно различным фонологическим школам, русский язык содержит от 39 до 43 фонем. Также под единицами речи могут пониматься аллофоны или их части. В настоящем описании под термином "аллофон" понимается конкретная реализация фонемы в речи, обусловленная ее фонетическим окружением. Аллофон, учитывающий по 1 фонеме перед и после данной, называют трифоном. Как правило, фонемы или трифоны моделируются скрытой марковской моделью с состояниями 1-3 (состояние 1 - вход в звук, переход с предыдущего, состояние 2 - стабильная часть, состояние 3 - выход из звука, переход в следующий), при этом состояния некоторых Трифонов "связываются" вместе, чтобы обеспечить достаточное количество данных для обучения редких Трифонов. Такие связанные состояния называют "сенонами", и именно им соответствуют выходы нейронной сети, т.е. нейронная сеть классифицирует векторы признаков речи на классы сенонов, оценивает вероятности каждого сенона при наблюдаемом векторе признаков.

В ходе экспериментов было выявлено, что в одном из вариантов реализации оптимальные результаты обеспечивает конфигурация глубокой нейронной сети, содержащая 6 скрытых слоев по 1536 нейронов с сигмоидами в каждом и выходной софтмакс-слой с 13000 нейронов, соответствующих сенонам акустической модели на основе гаусовых смесей. При этом оптимальная конфигурация зависит от объема обучающих данных.

Обучающую выборку формируют из фонограмм различных дикторов. Фонограммы могут быть получены любым известным способом, например путем записи телефонных переговоров. В данном варианте осуществления дикторы говорят на одном языке. Для каждой фонограммы из обучающей выборки заранее вычисляют низкоуровневые акустические признаки (мел-частотные кепстральные коэффициенты, например размерности 12, либо логарифмы энергии в мел-частотных полосах, например размерности 23). Под низкоуровневыми акустическими признаками понимаются признаки, извлекаемые напрямую из речевого сигнала или его спектра методами цифровой обработки сигналов. Они несут в себе важную информацию о сигнале, но являются трудно интерпретируемыми с точки зрения классификации единиц речи. При этом в других вариантах реализации на вход нейронной сети можно подавать другие низкоуровневые акустические признаки, например коэффициенты перцептивного линейного предсказания (perceptual linear prediction, PLP), энергии выходов банка гамматонных фильтров (gammatone filterbank, GTFB) и т.д. Предложенные низкоуровневые признаки мало отличаются по уровню информативности и могут быть использованы как по отдельности, так и в комбинации без ухудшения качества обучения нейронной сети.

Кроме того, из каждой фонограммы извлекают малоразмерное представление дикторской информации, содержащейся в фонограмме, в частности извлекают i-вектора, например размерности 50. Извлечение i-векторов проводят, например, с использованием универсальной фоновой модели (Universal Background Model, UBM), которая была обучена заранее. I-вектор аккумулирует в себе дикторскую информацию, и при этом в некоторых вариантах осуществления представляет собой малоразмерный вектор, кодирующий отклонение распределения акустических признаков фонограммы от распределения, оцененного по всей обучающей выборке. В других вариантах реализации, в которых требуется сравнительно меньшая точность обучения нейронных сетей, возможно извлечение дикторской информации в виде коэффициентов максимума правдоподобия линейной регрессии в пространстве признаков (feature space Maximum Likelihood Linear Regression, fMLLR).

На первом этапе глубокую нейронную сеть обучают предсказывать вероятности состояний сенонов, соответствующих отдельному кадру речи, с использованием только низкоуровневых акустических признаков по критерию минимума кросс-энтропии.

Кросс-энтропия показывает, насколько распределение вероятностей на выходе нейронной сети соответствует реально наблюдаемому на данном кадре сенону. Чем ближе вероятность данного сенона к единице, а остальных сенонов к нулю, тем кросс-энтропия на данном кадре будет ниже. Таким образом, кросс-энтропия является мерой средней точности классификации отдельных кадров речи по всей обучающей выборки, и чем она меньше, чем точнее данная нейронная сеть способна предсказывать сеноны. Иными словами, минимизация кросс-энтропии эквивалентна снижению средней покадровой ошибки классификации.

После того как обучение сошлось по критерию минимума кросс-энтропии, подают на вход глубокой нейронной сети исходные низкоуровневые акустически признаки, дополненные i-вектором, предварительно расширив входной слой глубокой нейронной сети на размерность дополнительных признаков путем дополнения матрицы слоя нулями, что позволит сохранить поведение сети за счет домножения нулей на компоненты i-вектора. Таким образом, на каждом кадре входной вектор состоит из 2 частей - первая часть (низкоуровневые акустические признаки) отличается от кадра к кадру, вторая (i-вектор) - одинакова для всех векторов одной фонограммы. При этом каждый голос диктора характеризуется набором особенностей, которые позволяют воспринимать его как голос именно этого диктора. Эти особенности можно трактовать как координаты в пространстве, поэтому каждый голос можно считать точкой в пространстве голосов, и если два голоса близки по каким-то параметрам, то соответственно точки также будут находиться близко в пространстве голосов и соответствующие им i-векторы также будут близко в пространстве голосов. Таким образом, за счет расширения входных векторов признаков i-вектором, характеризующим "расположение голоса диктора в пространстве голосов", обеспечивается распознавание речи различных (произвольных) дикторов. Это объясняется тем, что, поскольку в обучающей выборке дикторов обычно много, сеть приобретает способность использовать информацию о том, из какой области пространства голосов поступил входной i-вектор. Таким образом, во время распознавания произвольного диктора его i-вектор окажется в области пространства, где были i-векторы дикторов из обучающей выборки, благодаря чему нейронная сеть сможет с максимальной эффективностью учитывать эту информацию; другими словами, нейронная сеть уже будет представлять, как эту информацию следует обработать.

Обученную с использованием только низкоуровневых акустических признаков глубокую нейронную сеть дообучают по критерию минимума суммы кросс-энтропии, который позволяет комбинировать все величины, для одновременного их снижения, и дополнительного регуляризирующего слагаемого, которое контролирует отклонение весов обучаемой таким образом глубокой нейронной сети от весов глубокой нейронной сети, обученной с использованием только низкоуровневых акустических признаков, что позволяет избежать сильного изменения весов глубокой нейронной сети по сравнению с хорошим (качественным) начальным приближением.

Важно отметить, что минимизация кросс-энтропии эквивалентна снижению средней покадровой ошибки классификации (Frame Error Rate, FER), а целью распознавания речи является не получение результатов классификации отдельных кадров, как в случае использования критерия минимума кросс-

энтропии, а получение последовательности произнесенных слов. И мерой ошибки системы распознавания является пословная ошибка (Word Error Rate, WER). Безусловно, пословная ошибка и покадровая ошибка сильно коррелируют, и снижение покадровой ошибки до нуля практически неизбежно ведет к идеально точному распознаванию (при условии использования качественного лексикона и языковой модели). Однако на практике снижение до нуля покадровой ошибки недостижимо. Пословную ошибку исключительно сложно использовать в качестве критерия обучения нейронной сети, т.к. она является не дифференцируемой (по параметрам сети) и трудно вычислимой в ходе обучения. По этой причине используют другие критерии обучения, в частности последовательно-дискриминативные, косвенно направленные именно на уменьшение пословной ошибки, но более доступные с вычислительной точки зрения. Эти критерии рассматривают лучшую гипотезу о последовательности распознанных слов в декодере и стремятся таким образом скорректировать параметры нейронной сети, чтобы одновременно приблизить ее к истинной последовательности слов и максимально отдалить от всех "конкурирующих" гипотез. Критерий минимума среднего риска, вычисляемого по состояниям (state-level Minimum Bayes Risk, sMBR), - лишь один из ряда известных критериев этого класса. Он показывает сравнимую с остальными подобными критериями точность, однако является более легким с вычислительной точки зрения. Таким образом, после дообучения глубокой нейронной сети по критерию минимума суммы кросс-энтропии и дополнительного регуляризирующего слагаемого ее дообучают по критерию минимума среднего риска, что дает существенный прирост в точности обучения нейронной сети.

После того как обучение сошлось, матрицу весов последнего скрытого слоя обученной сети подвергают сингулярному разложению и снижают ее ранг путем отбрасывания наименьших сингулярных чисел. В результате такой операции последний слой исходной сети оказывается заменен на 2 слоя, один из которых - линейный и содержит меньше нейронов по сравнению с входным слоем. Этот слой называют слоем "узкого горла" (bottleneck), или малоразмерным слоем. Часть информации при прохождении через малоразмерный слой необратимо теряется, но в результате сохраняются наиболее существенные ее составляющие. Первоначальное обучение без малоразмерного слоя позволяет привести веса остальных слоев к значениям, достаточно близким к оптимальным, что облегчает дообучение сети после внедрения малоразмерного слоя, т.е. последовательное обучение сети сначала без малоразмерного слоя, а потом с ним позволяет двигаться путем последовательных улучшений, т.е. последовательной настройкой параметров (весов). Экспериментально было выяснено, что обучение нейронной сети, изначально имеющей малоразмерный слой, снижает качество и повышает сложность ее обучения.

В результате предыдущего обучения выходы глубокой нейронной сети имеют хорошие (качественные) распределения вероятностей сенонов, которые уже настроены по последовательно-дискриминативному критерию. Поскольку в результате сингулярного разложения матрица весов последнего слоя претерпела изменения, полученная глубокая нейронная сеть уже не является оптимальной с точки зрения критерия предыдущего этапа обучения. Поэтому глубокую нейронную сеть теперь уже с малоразмерным слоем еще раз дообучают, используя распределения из предыдущего обучения в качестве целевых распределений. При этом дообучение нейронной сети происходит по использованному уже ранее критерию минимума кросс-энтропии до сходимости, что позволяет улучшить качество извлекаемых высокоуровневых малоразмерных признаков из малоразмерного слоя. Высокоуровневость признаков обусловлена тем, что глубокая нейронная сеть с малоразмерным слоем, обученная по критерию минимума кросс-энтропии, способна обеспечивать почти столь же низкие значения кросс-энтропии, что и глубокая нейронная сеть без малоразмерного слоя, обученная по тому же критерию. Таким образом, признаки, извлеченные с выходов малоразмерного слоя, содержат в себе всю существенную информацию из речевого сигнала, содержащуюся в исходных низкоуровневых акустических признаках и i-векторе.

Кроме того, после того как глубокая нейронная сеть обучена до сходимости, слои нейронной сети, находящиеся после малоразмерного слоя, могут быть удалены, что позволит обученной глубокой нейронной сети стать "экстрактором" новых дикторозависимых малоразмерных высокоуровневых признаков, т.е. при подаче на вход нейронной сети вектора низкоуровневых признаков, расширенных (дополненных) i-вектором, как было описано ранее, на выходе могут быть получены значения активации малоразмерного слоя (слоя узкого горла), которые являются малоразмерным, дикторозависимым и высокоуровневым представлением.

Предложенный способ может быть применен для получения многоязычных дикторозависимых малоразмерных высокоуровневых акустических признаков речи. Для этого на вход нейронной сети подают низкоуровневые речевые признаки по меньшей мере двух различных языков и соответствующую им дикторскую информацию (i-вектор), при этом данные различных языков на вход нейронной сети подают вперемежку в случайном порядке. В данном случае архитектура нейронной сети должна быть предназначена для многозадачного обучения, т.е. нейронная сеть должна иметь несколько скрытых слоев, веса которых будут являться общими для данных из обучающего множества на всех языках, содержащих низкоуровневые речевые признаки и дикторскую информацию, и множество выходных слоев, каждый из которых обрабатывает данные на одном из указанных по меньшей мере двух языков. Таким образом, при обучении с использованием двух языков, например, если на вход нейронной сети подают данные, относящиеся к первому языку, то после прохождения скрытых слоев данные попадают на первый выходной

слой, относящийся непосредственно к первому языку, где вычисляется ошибка, которая методом обратного распространения корректирует общие для двух языков веса скрытых слоев нейронной сети. Далее, если на вход нейронной сети подают данные, относящиеся ко второму языку, то они по тому же принципу попадают на соответствующий им второй выходной слой, где также вычисляется ошибка, с помощью которой также корректируют общие для двух языков веса скрытых слоев нейронной сети. Таким образом нейронная сеть обучается по данным на всех имеющихся языках. При этом процесс обучения нейронной сети аналогичен описанному выше в отношении одного языка, а по завершении обучения с выхода малоразмерного слоя извлекают многоязычные дикторозависимые малоразмерные признаки, которые представляют собой высокоуровневые признаки, содержащие в себе информацию, относящуюся ко всем языкам обучающей выборки, и, как следствие, устойчивые к изменению языка при распознавании речи. При этом обучение одной многоязычной акустической модели нейронной сети может потребовать меньше вычислений, чем обучение нескольких многоязычных акустических моделей для каждого языка в отдельности. Кроме того, при ограниченности данных того или иного языка, когда соответствующие данные для обучения недоступны или дорогостоящие в получении, многоязычная акустическая модель может предложить лучшую точность в сравнении с одноязычными акустическими моделями, полученными с использованием ограниченных данных соответствующего языка.

Экспериментально было выявлено, что именно предложенный порядок действий при обучении глубокой нейронной сети является наиболее подходящим для получения дикторозависимых малоразмерных высокоуровневых признаков, обладающих высокой информативностью и позволяющих обеспечить адаптацию акустической модели к акустической вариативности речевого сигнала и, как следствие, высокую точность распознавания речи такой моделью.

Высокоуровневые признаки, извлекаемые с выхода малоразмерного слоя обученной нейронной сети, впоследствии могут быть использованы для обучения другой нейронной сети для распознавания речи.

На фиг. 3 изображено обучение другой нейронной сети Б для распознавания речи, обозначенной как блок Б (левая часть схемы), на входной слой 4 которой поступают высокоуровневые признаки с малоразмерного слоя 2а обученной нейронной сети А, обученной предложенным способом и обозначенной как блок А (левая часть схемы). На вход нейронной сети Б поступает вектор, являющийся объединением векторов с текущего кадра (задержка 0), а также с кадров, находящихся за 5, 10 и 15 кадров до текущего и через 5, 10, 15 кадров после текущего. В результате, при размерности малоразмерных признаков, например, 100, на вход второй сети Б поступает вектор размерностью 700. Нейронная сеть Б, которую обучают для распознавания речи, содержит входной слой 4, который принимает указанный вектор, скрытые слои 5, количество которых выбирается экспериментально, и выходной слой 6, являющийся выходом нейронной сети Б.

В таблице приведено сравнение значений пословной ошибки распознавания (WER) глубоких нейронных сетей, обученных на дикторозависимых малоразмерных высокоуровневых признаках, полученных предложенным способом (speaker dependent bottleneck features - Deep Neural Network, SDBN-DNN), и глубоких нейронных сетей, обученных дикторoадаптивным способом с использованием i-векторов (Deep Neural Network - i-vector, DNN-ivec). Из таблицы видно, что использование SDBN-признаков обеспечивает снижение ошибки распознавания. При этом обучение глубокой нейронной сети по критерию минимума среднего риска (state-level Minimum Bayes Risk, sMBR) обеспечивает более низкую ошибку распознавания в сравнении с обучением глубокой нейронной сети только по критерию минимума кросс-энтропии (Cross-Entropy, CE).

Результаты распознавания речи

Акустическая модель	Критерий обучения	WER, %
DNN-ivec	CE	23,8
SDBN-DNN	CE	22,0
DNN-ivec	sMBR	21,7
SDBN-DNN	sMBR	19,5

Настоящее изобретение не ограничено конкретными вариантами реализации, раскрытыми в описании в иллюстративных целях, и охватывает все возможные модификации и альтернативы, входящие в объем настоящего изобретения, определенный формулой изобретения.

ФОРМУЛА ИЗОБРЕТЕНИЯ

1. Способ получения малоразмерных высокоуровневых акустических признаков речи, согласно которому

обеспечивают наличие низкоуровневых признаков речи и соответствующей им дикторской информации;

обучают нейронную сеть с использованием низкоуровневых признаков речи;

дообучают нейронную сеть с использованием низкоуровневых признаков речи, дополненных дикторской информацией;

вводят малоразмерный слой в состав нейронной сети;

дообучают нейронную сеть с малоразмерным слоем с использованием низкоуровневых признаков речи, дополненных дикторской информацией;

извлекают с выхода малоразмерного слоя нейронной сети малоразмерные высокоуровневые акустические признаки речи.

2. Способ по п.1, согласно которому после обучения нейронной сети с использованием низкоуровневых речевых признаков ее входной слой расширяют путем дополнения матрицы входного слоя нулевыми столбцами.

3. Способ по любому из пп.1 и 2, согласно которому низкоуровневые речевые признаки имеют вид мел-частотных кепстральных коэффициентов.

4. Способ по любому из пп.1 и 2, согласно которому низкоуровневые речевые признаки имеют вид логарифмов энергии в мел-частотных полосах.

5. Способ по любому из пп.1-4, согласно которому дикторская информация имеет вид малоразмерного i-вектора.

6. Способ по любому из пп.1-5, согласно которому обучение нейронной сети с использованием низкоуровневых речевых признаков проводят по критерию минимума кросс-энтропии.

7. Способ по любому из пп.1-6, согласно которому дообучают нейронную сеть с использованием низкоуровневых речевых признаков, дополненных дикторской информацией, по критерию минимума суммы кросс-энтропии и дополнительного регуляризирующего слагаемого.

8. Способ по п.7, согласно которому дообучают нейронную сеть с использованием низкоуровневых речевых признаков, дополненных дикторской информацией, по последовательно-дискриминативному критерию.

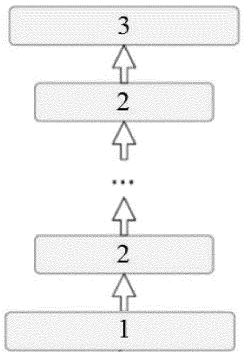
9. Способ по любому из пп.1-8, согласно которому вводят малоразмерный слой путем низкоранговой факторизации матрицы весов последнего скрытого слоя.

10. Способ по п.9, согласно которому низкоранговую факторизацию матрицы весов последнего скрытого слоя обеспечивают сингулярным разложением.

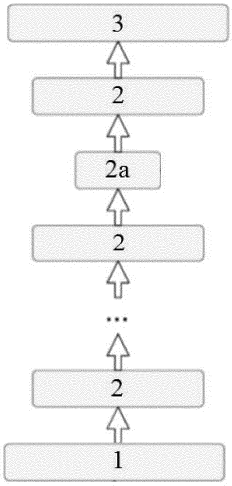
11. Способ по любому из пп.1-10, согласно которому после завершения дообучения нейронной сети с малоразмерным слоем слои, расположенные после малоразмерного слоя нейронной сети, удаляют.

12. Способ по любому из пп.1-11, согласно которому подают низкоуровневые речевые признаки по меньшей мере двух различных языков и соответствующую им дикторскую информацию на вход нейронной сети и извлекают с выхода малоразмерного слоя нейронной сети многоязычные малоразмерные высокоуровневые акустические признаки речи.

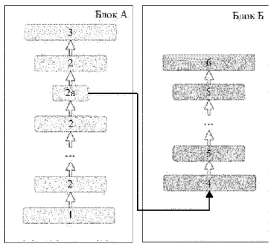
13. Способ по п.12, согласно которому количество выходных слоев нейронной сети равно количеству языков, при этом веса каждого из выходных слоев настраивают только по данным соответствующего языка, а веса всех скрытых слоев настраивают по данным всех из указанных по меньшей мере двух языков.



Фиг. 1



Фиг. 2



Фиг. 3