(19) Евразийское патентное ведомство

(11) 045617 (13) B1

(12) ОПИСАНИЕ ИЗОБРЕТЕНИЯ К ЕВРАЗИЙСКОМУ ПАТЕНТУ

(45) Дата публикации и выдачи патента
2023.12.12

(21) Номер заявки
202392106

(22) Дата подачи заявки
2021.07.26

(51) Int. Cl. ***G10L 15/16*** (2006.01)
G10L 25/63 (2013.01)
G06N 3/08 (2006.01)

(54) ОБУЧЕНИЕ НЕЙРОННОЙ СЕТИ И СЕГМЕНТАЦИЯ АУДИОЗАПИСИ ДЛЯ РАСПОЗНАВАНИЯ ЭМОЦИИ

(43) **2023.09.27**
(86) **PCT/RU2021/000316**
(87) **WO 2023/009020 2023.02.02**

(71)(73) Заявитель и патентовладелец:
ОБЩЕСТВО С ОГРАНИЧЕННОЙ ОТВЕТСТВЕННОСТЬЮ "ЦРТ-ИННОВАЦИИ" (RU)

(72) Изобретатель:
Тимофеев Илья Валерьевич, Агафонов Юрий Олегович, Акулов Артем Викторович (RU)

(74) Представитель:
Нилова М.И. (RU)

(56) US-A1-20210192332
KANNAN VENKATARAMANAN et al.: Emotion Recognition from Speech. 22.12.2019 [retrieved on 2022-03-28]. Retrieved from https://arxiv.org/pdf/1912.10458.pdf
DIAS ISSA et al.: Speech emotion recognition with deep convolutional neural networks. Biomedical Signal Processing and Control 59 (2020) 101894, 27.02.2020 [retrieved on 2022-03-28]. Retrieved from https://reader.elsevier.com/reader/sd/pii/S1746809420300501?token=74F6416D88965B5EB7522FD1A408A92CE9F4EB5FA4F477EA683A815AAC5888AD6E5C62719C9D11D50ABB0454051A543A&originRegion=eu-west-1 &originCreation=20220328070107
SHUIYANG MAO et al.: Emotion Profile Refinery for Speech Emotion Classification. 12.08.2020 [retrieved on 2022-03-28]. Retrieved from https://arxiv.org/pdf/2008.05259.pdf
SHANSHAN WANG et al. Audio-Visual Scene Classification: Analysis of DCASE 2021 Challenge Submissions. 28.05.2021 [retrieved on 2022-03-28]. Retrieved from https://arxiv.org/pdf/2105.13675vl.pdf
VLADIMIR CHERNYKH et al.: Emotion Recognition From Speech With Recurrent Neural Networks. 27.01.2017 [retrieved on 2022-03-28]. Retrieved from https: // arxiv.org/pdf/1701.08071vl.pdf

(57) Изобретение относится к способу обучения нейронной сети для задачи распознавания эмоций в сегментах речи и системе для сегментации речи и распознавания эмоции в указанных сегментах речи, в частности изобретение направлено на выделение сегментов речи с необходимой эмоцией из длительных аудиозаписей. Предложенный способ обучения нейронной сети для задачи распознавания эмоции в сегменте речи, включает следующие этапы: замораживают сверточную нейронную сеть OpenL3; формируют базу размеченных реплик, содержащую реплики не более 10 с, к каждой из которых с использованием разметчиков присвоена соответствующая эмоциональная метка или метка шума, а разметчики представляют собой группу разметчиков, из которой исключены разметчики, не соответствующие уровню согласованности 0,4 по каппе Флейса; обучают рекуррентную нейронную сеть малой емкости, построенную на указанной предварительно обученной сверточной нейронной сети OpenL3, с использованием сформированной базы размеченных реплик; размораживают верхние слои указанной предварительно обученной сверточной нейронной сети OpenL3 для прохождения дообучения нейронной сети.

045617 B1

045617 B1

Область техники

Настоящее изобретение относится к способу обучения нейронной сети для задачи распознавания эмоций в сегментах речи и системе для сегментации аудиозаписи и распознавания эмоции в указанных сегментах речи, в частности изобретение направлено на выделение сегментов речи с необходимой эмоцией из длительных аудиозаписей.

Уровень техники

Учёт эмоционального состояния клиента в сценариях работы голосовых ассистентов, в частности эмпатическая (адекватная) реакция голосового робота на эмоции клиента (собеседника), и своевременное решение спорных ситуаций играет все более важную роль в Интернет-индустрии.

Наиболее распространенными методами моделирования и классификации в области распознавания речевых эмоций являются смеси гауссовских распределений (Gaussian Mixture Models, GMM), скрытые марковские модели (Hidden Markov Models, HMM), метод опорных векторов (Support Vector Machines, SVM) и глубокие нейронные сети (Deep Neural Networks, DNN).

Одним из главных критериев успешного обучения глубокой нейронной сети для задачи распознавания эмоций по речи является наличие большого объема примеров. Однако, поскольку эмоциональные речевые данные трудно получить и сложно маркировать, объем баз данных, содержащих размеченную речь, часто ограничен.

Люди проявляют эмоции в речи значительно реже, чем в мимике лица, это обусловлено культурными нормами. Как результат, подавляющее большинство обработанного материала со спонтанной речью не содержит эмоций. Кроме того, проявление спонтанных эмоций в голосе значительно отличается от наигранных "актерских" эмоций, по этой причине обученные на "актерских" эмоциях нейронные сети будут иметь низкую точность.

Сложность маркировки речевых данных заключается в том, что метки к каждой эмоции выставляются вручную и разметчики достаточно индивидуально интерпретируют проявление эмоций в речи, т.е. одно и тоже проявление эмоций разными разметчиками часто интерпретируются по-разному. Таким образом, большая часть данных, содержащих размеченную речь, не имеют однозначной разметки, поскольку мнения разметчиков зачастую расходятся.

Указанные проблемы получения базы данных приводят к недостаточному обучению глубоких нейронных сетей и, соответственно, к низкой точности оценки эмоций в аудиозаписи.

Кроме того, решение спорных ситуаций, возникающих в общении клиента с голосовым ассистентом, требуют много времени на анализ, поскольку, как правило, результатом распознавания эмоций в речи является присвоение эмоции всей аудиозаписи.

Таким образом, существует необходимость в создании решения, которое бы обеспечивало распознавание эмоций в речи с высокой точностью и прослушивание спорных сессий в кратчайшие сроки для эффективной работы с клиентом.

Раскрытие сущности изобретения

Согласно первому аспекту настоящего изобретения предложен способ обучения нейронной сети для задачи распознавания эмоции в сегменте речи, согласно которому замораживают сверточную нейронную сеть OpenL3, предварительно обученную на большом неразмеченном корпусе данных в режиме самообучения; формируют базу размеченных реплик, содержащую реплики не более 10 с, к каждой из которых с использованием разметчиков присвоена соответствующая эмоциональная метка или метка шума, причем для реплик, к которым большинством указанных разметчиков присвоена одинаковая эмоциональная метка, используется техника жесткой метки, для остальных реплик используется техника мягкой метки, а разметчики представляют собой группу разметчиков, из которой исключены разметчики, не соответствующие уровню согласованности 0,4 по каппе Флейса; обучают рекуррентную нейронную сеть малой емкости, построенную на указанной предварительно обученной сверточной нейронной сети OpenL3, с использованием сформированной базы размеченных реплик, причем реплики с жесткими метками, использующие функцию потерь Cross Entropy, и мягкими метками, использующие функцию потерь mean squared error (MSE), передают пакетами поочередно; размораживают верхние слои указанной предварительно обученной сверточной нейронной сети OpenL3 для прохождения дообучения нейронной сети.

Предлагаемый способ обучения нейронной сети позволяет получить нейронную сеть способную распознавать эмоцию для каждого сегмента речи, выделенного из аудиозаписи.

Согласно одному из вариантов реализации техника жесткой метки используется для реплик, к которым присвоена одинаковая эмоциональная метка 80% разметчиков.

Согласно еще одному из вариантов реализации эмоциональная метка отражает злость, грусть, радость или нейтральность.

Согласно второму аспекту настоящего изобретения предложена система сегментации аудиозаписи и распознавания эмоции в сегменте речи, содержащая блок выделения речевой активности (Voice Activity Detector, VAD), выполненный с возможностью выделения реплик из аудиозаписи, блок деления реплик, выполненный с возможностью деления реплик на фрагменты по 3 с, блок распознавания эмоций, содержащий нейронную сеть, обученную согласно способу по п.1, которая содержит блок получения

мел-спектрограммы, выполненный с возможностью получения мел-спектрограммы из фрагментов, блок сверточной нейронной сети OpenL3, выполненный с возможностью преобразования мел-спектрограммы в последовательность векторов размерности 512, и блок рекуррентной нейронной сети малой емкости, выполненный с возможностью формирования из полученной последовательности векторов вектора вероятностей наличия в каждом фрагменте соответствующей эмоции или шума. Система также содержит блок фильтрации, выполненный с возможностью определения величины вероятности наличия соответствующей эмоции в каждом фрагменте, по значению которой фильтруют каждый фрагмент, содержащий соответствующую эмоцию, с использованием порогового значения для выявления яркости эмоции и объединения последовательных фрагментов с одинаковой эмоцией в сегменты.

Предлагаемая система позволяет сегментировать аудиозапись на короткие сегменты речи и распознавать эмоцию для каждого выделенного сегмента речи.

Согласно одному из вариантов реализации одно пороговое значение используют для разных эмоций. Согласно другому из вариантов реализации для каждой эмоции используют соответствующее пороговое значение.

Согласно еще одному из вариантов реализации каждый сегмент, содержащий соответствующую эмоцию, имеет информацию о времени его начала и конца в аудиозаписи.

Краткое описание чертежей

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

фиг. 1 - функциональная схема способа обучения нейронной сети для задачи распознавания эмоций в сегментах речи;

фиг. 2 - функциональная схема формирования базы размеченных реплик;

фиг. 3 - архитектура обученной нейронной сети для задачи распознавания эмоции в сегменте речи;

фиг. 4 - система для сегментации речи и распознавания эмоции в указанных сегментах речи.

Подробное описание

Способ обучения нейронной сети для задачи распознавания эмоции в сегменте речи, в соответствии с различными вариантами реализации настоящего изобретения, 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.

На фиг. 1 представлена схема 10, опираясь на которою может быть реализован способ обучения нейронной сети для задачи распознавания эмоции в сегменте речи (далее - способ обучения нейронной сети) в соответствии с одним из вариантов осуществления настоящего изобретения.

Для обучения нейронной сети необходимо наличие обучающей базы, а именно базы размеченных реплик (фиг. 1, блок 100). Схема формирования базы размеченных реплик показана на фиг. 2. В качестве исходных данных для формирования базы размеченных реплик используют аудиозаписи колцентра, содержащие спонтанных эмоции в диалоге (фиг. 2, блок 101). Также, могут использоваться любые другие аудиозаписи, содержащие эмоциональную спонтанную речь, например аудиозаписи интервью с разными людьми и аудиозаписи ток-шоу. Использование аудиозаписей, содержащих "актерские" (наигранные) эмоции, для обучения нейронной сети приведет к снижению точности распознавания эмоций.

В традиционных моделях распознавания эмоций используют всю аудиозапись или аудиозапись, разделенную на достаточно длинные реплики, поскольку такие аудиозаписи или реплики содержат большое количество характеристик, позволяющих выявить соответствующую эмоцию. Таким образом, указанные модели выявляют эмоцию всей аудиозаписи в целом.

В предлагаемом способе используют аудиозапись, разделенную на более короткие реплики фиксированной длины. В частности, предлагаемый способ обучения нейронной сети направлен на распознавание эмоции каждого сегмента аудиозаписи индивидуально. Указанная задача распознавания эмоций в коротких сегментах речи является более сложной по сравнению с традиционными методами, поскольку характеристик в коротком сегменте, позволяющих выявить соответствующую эмоцию, меньше, но преимуществом является то, что появляется возможность более точно вычленять эмоции в ситуациях, когда разные эмоции смешаны между собой. В предпочтительном варианте аудиозаписи, содержащие эмоциональную спонтанную речь, разделяют на реплики длиной не более 10 с, с использованием модели выделения речевой активности (Voice Activity Detector, (VAD)) (фиг. 2, блок 102). В дальнейшем из указанных реплик используется только 3-секундный фрагмент речи, выбранный произвольно.

Как было указано ранее, используемые аудиозаписи содержат спонтанные речевые взаимодействия, в которых границы между эмоциями (эмоциональными классами) размыты и смешаны. Таким образом, чтобы обеспечить высокую точность распознавания эмоции каждой реплики (сегмента реплики), важно обеспечить согласованность разметчиков, присваивающих соответствующую эмоциональную метку реплике. В предпочтительном варианте присваивание репликам соответствующих эмоциональных меток выполняют не менее 7 разметчиков (фиг. 2, блок 103). Для блока реплик (не менее 100) проводят выбор не менее 5 наиболее согласованных разметчиков по каппе Флейса, добиваясь уровня согласованности 0,4, что позволяет исключить разметчиков, чье представление о проявлении эмоций в речи отличается от

большинства (фиг. 2, блок 104).

Помимо эмоциональной метки, которая отражает злость, грусть, радость или нейтральность, разметчики также присваивают соответствующей реплике метку шума. Метка шума нужна для обеспечения устойчивости к ошибкам VAD. В частности, при разделении аудиозаписи на речевые реплики VAD может ошибаться, принимая шумовые помехи (шум микрофона, ветер и др.) за речь. Таким образом, если при обучении нейронной сети не учитывать шум, в дальнейшем при ее использовании она будет выдавать большой процент ошибок. Предлагаемый способ обучения нейронной сети предполагает возможность наличия шума среди реплик, что позволяет минимизировать ошибку нейронной сети, а именно предугадать присваивание реплике с шумом эмоциональной метки, например злости, которая схожа по характеристикам с ним.

Значение выраженности эмоции (злость, грусть, радость, нейтральность) в реплике определяют долей разметчиков, отдавших свой голос за указанную эмоцию (фиг. 2, блок 105). В частности, для реплик, к которым большинством указанных разметчиков присвоена одинаковая эмоциональная метка, используется техника жесткой метки. В предпочтительном варианте техника жесткой метки используется для реплик, к которым присвоена одинаковая эмоциональная метка 80% разметчиков. Для остальных неоднозначно размеченных реплик используется техника мягкой метки, которая назначает силу эмоциональной метки пропорционально доли разметчиков, отнесших реплику к соответствующей эмоциональной метке. Реплики категории шум отмечают отдельно только категорией шум.

Обучаемая нейронная сеть является глубокой нейронной сетью, использующей технику переноса знаний (transfer learning), а именно между сверточной нейронной сетью (convolutional neural network, CNN), OpenL3, предварительно обученной на большом неразмеченном корпусе данных в режиме самообучения, и рекуррентной нейронной сетью малой емкости, следующей за указанной нейронной сетью OpenL3.

Обучение нейронной сети выполняют в два этапа. На первом этапе (фиг. 1, блок 200) сверточную нейронную сеть OpenL3 замораживают (веса не обновляют) и обучают только рекуррентную нейронную сеть малой емкости, построенную на указанной сверточной нейронной сети OpenL3. Веса рекуррентной нейронной сети инициализируют случайным образом. Рекуррентную нейронную сеть обучают с использованием сформированной базы размеченных реплик, причем реплики с жесткими метками и мягкими метками передаются пакетами поочередно и с использованием функций потерь Cross Entropy и mean squared error (MSE) соответственно. При обучении используется оптимизатор Adam. Обучение продолжается до стабилизации функции потерь на валидации.

На втором этапе (фиг. 1, блок 300) размораживают верхние слои указанной предварительно обученной сверточной нейронной сети OpenL3, которые со следующей за ними рекуррентной нейронной сетью проходят ограниченное количество итераций обучения.

Архитектура нейронной сети, обученной для задачи распознавания эмоции в сегменте речи, показана на фиг. 3. Блок MelSpec содержит блоки FFT and MEL и выполнен с возможностью получения мел-спектрограммы из фрагментов речи (быстрое преобразование Фурье (FFT) для перехода в частотную область и последующая агрегация частот с учетом чувствительности человеческого уха), не требует обучения. Блок OpenL3 содержит CNN Blockl (1,64), CNN Block2 (64,128), CNN Block3, (128,256) и CNN Block4 (256,512), причем CNN Block содержит следующие функции в указанной полседовательности: Conv2d (in, out), BatchNorm2d, Relu, Conv2d (out, out), BatchNormld, Relu и MaxPool2d. Использование блока OpenL3, предоставляющего собой сверточную нейронную сеть, предварительно обученную на миллионах аудиосегментов, позволяет решить проблему предварительной обработки звука, в частности в предлагаемом способе выступает в качестве экстрактора признаков, преобразующего мел-спектрограмму в последовательность векторов размерности 512. Стоит отметить, что ранее сверточная нейронная сеть OpenL3 применялась только для детектирования акустических событий, анализа музыки или видео.

Последовательность векторов, отражающих состояние звука фрагмента за секунду с перекрытием пол-секунды, предают в качестве токенов на блок Emo, предоставляющего собой рекурентную нейронную сеть, а именно long short-term memory (LSTM). Блок Emo содержит следующие функции в указанной последовательности: BathNormld, LSTM, ReLu, Linear, ReLu и Linear Прохождение токенов через два Dence блока с активацией ReLu формируется вектор вероятностей наличия следующих эмоций во фрагменте: злость, грусть, радость, нейтральность или шум. Добавление типа шум позволяет сократить ложные срабатывания в случае специфичных микрофонных помех.

На фиг. 4 представлена система 400 сегментации аудиозаписи и распознавания эмоции в сегменте речи (далее - система) в соответствии с одним из вариантов осуществления настоящего изобретения.

Согласно системе, выделяют реплики с использованием блока выделения речевой активности (Voice Activity Detector, VAD) (фиг. 4, блок 401), а с использованием блока деления реплик (фиг. 4, блок 402) разделяют реплики на фрагменты по 3 с.

Далее в блоке распознавания эмоций (фиг. 4, блок 403) определяют вероятности эмоционального состояния для каждого фрагмента с использованием нейронной сети, обученной предложенным выше способом. В частности, передают полученные фрагменты в блок получения мел-спектрограммы, выполненный с возможностью получения мел-спектрограммы из фрагментов для дальнейшего ее преобразова-

ния в последовательность векторов размерности 512 с использованием блока сверточной нейронной сети OpenL3, и формирования из полученной последовательности векторов вектора вероятностей наличия в каждом фрагменте соответствующей эмоции или шума с использованием блока рекуррентной нейронной сети малой емкости.

После чего в блоке фильтрации (фиг. 4, блок 404) определяют величину вероятности наличия соответствующей эмоции в каждом фрагменте, по значению которой фильтруют каждый фрагмент, содержащий соответствующую эмоцию, с использованием порогового значения для выявления яркости эмоции. Другими словами, пороговое значение влияет на точность и полноту (Precision и Recall) эмоций, содержащихся в полученных фрагментах, таким образом для каждого, из полученных фрагментов, система определяет меру уверенности в результате (Confidence), по значению которого выполняется фильтрация.

Пороговое значение подбирается в интервале от 0 до 1 для каждой задачи сегментации и может выставляться как единое для всех классов эмоций, так и специальное для каждой отдельной эмоции, таким образом, пользователь имеет возможность регулировать результат, выдаваемый системой.

Полученные последовательные фрагменты, содержащие одинаковую эмоцию, объединяют в сегменты, например могут быть объединены идущие подряд фрагменты, содержащие злость. Одиночные фрагменты также преобразуются в сегменты.

Так, например, при выставлении единого порогового значения, близкого к 0, пользователь получит большее количество сегментов для анализа, что снижает вероятность пропустить сегмент с необходимой яркостью (выраженностью) эмоции. При выборе порогового значения, близкого к 1, нейронная сеть будет определять сегменты более точно (с более яркими эмоциями), пользователь получит меньшее количество сегментов, что ускорит их анализ, но часть сегментов будет пропущена.

При выставлении порогового значения, специального для каждой отдельной эмоции, пользователь получит возможность отфильтровать сегменты с необходимой эмоцией.

Таким образом, предлагаемая система позволяет пользователю прослушивать не всю аудиозапись полностью, а только выделенные сегменты, содержащие необходимую яркость конретно определенной эмоции.

Предлагаемая система позволяет обеспечить высокую точность распознавания эмоций в речи и ускоренное прослушивание спорных сессий общения, например клиента и голосового ассистента, что делает возможным контролировать качество работы колцентра, в частности позволяет корректно определять проблему сессии и своевременно ее разрешать тем самым, благоприятно влияя на впечатление клиента о компании.

Кроме того, предлагаемое изобретение может быть использовано для автоматического составления подборки ярких моментов аудио или видео передачи, тем самым сокращая время на просмотр.

Также, поскольку каждый сегмент, содержащий соответствующую эмоцию, имеет информацию о времени его начала и конца в аудиозаписи, обеспечивается возможность делать разметку аудиозаписи по эмоциям.

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 обучения нейронной сети для задачи распознавания эмоции в сегменте речи, согласно которому

замораживают сверточную нейронную сеть OpenL3, предварительно обученную на большом неразмеченном корпусе данных в режиме самообучения,

формируют базу размеченных реплик, содержащую реплики не более 10 с, к каждой из которых с использованием разметчиков присвоена соответствующая эмоциональная метка или метка шума,

причем для реплик, к которым большинством указанных разметчиков присвоена одинаковая эмоциональная метка, используется техника жесткой метки, для остальных реплик используется техника мягкой метки,

причем разметчики представляют собой группу разметчиков, из которой исключены разметчики, не соответствующие уровню согласованности 0,4 по каппе Флейса,

обучают рекуррентную нейронную сеть малой емкости, построенную на указанной предварительно обученной сверточной нейронной сети OpenL3, с использованием сформированной базы размеченных реплик,

причем реплики с жесткими метками, использующие функцию потерь Cross Entropy, и мягкими метками, использующие функцию потерь mean squared error (MSE), передают пакетами поочередно, размораживают верхние слои указанной предварительно обученной сверточной нейронной сети OpenL3 для прохождения дообучения нейронной сети.

2. Способ по п.1, в котором техника жесткой метки используется для реплик, к которым присвоена

одинаковая эмоциональная метка 80% разметчиков.

3. Способ по п.1, в котором эмоциональная метка отражает злость, грусть, радость или нейтральность.

4. Система для сегментации аудиозаписи и распознавания эмоции в сегменте речи, содержащая

блок выделения речевой активности (Voice Activity Detector, VAD), выполненный с возможностью выделения реплик из аудиозаписи,

блок деления реплик, выполненный с возможностью деления реплик на фрагменты по 3 с,

блок распознавания эмоций, содержащий нейронную сеть, обученную согласно способу по п.1, которая содержит

блок получения мел-спектрограммы, выполненный с возможностью получения мел-спектрограммы из фрагментов,

блок сверточной нейронной сети OpenL3, выполненный с возможностью преобразования мел-спектрограммы в последовательность векторов размерности 512, и

блок рекуррентной нейронной сети малой емкости, выполненный с возможностью формирования из полученной последовательности векторов вектора вероятностей наличия в каждом фрагменте соответствующей эмоции или шума,

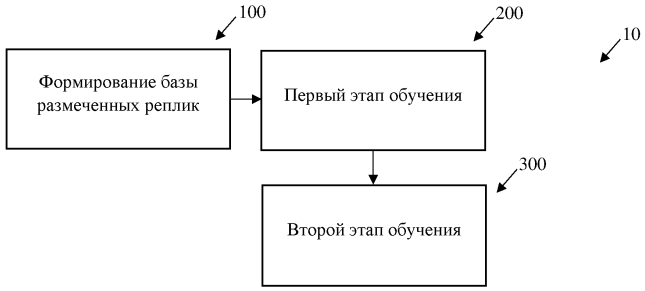
блок фильтрации, выполненный с возможностью определения величины вероятности наличия соответствующей эмоции в каждом фрагменте, по значению которой фильтруют каждый фрагмент, содержащий соответствующую эмоцию, с использованием порогового значения для выявления яркости эмоции и объединения последовательных фрагментов с одинаковой эмоцией в сегменты.

5. Система по п.1, в которой одно пороговое значение используют для разных эмоций.

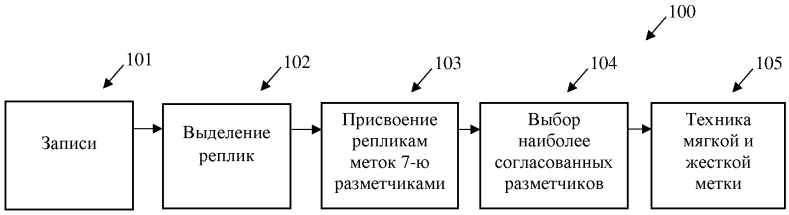
6. Система по п.1, в которой для каждой эмоции используют соответствующее пороговое значение.

7. Система по п.1, в которой каждый сегмент, содержащий соответствующую эмоцию, имеет информацию о времени его начала и конца в аудиозаписи.

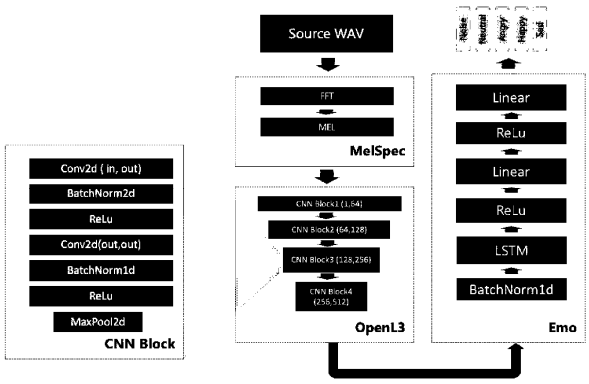
Обучение нейронной сети и сегментация аудиозаписи для распознавания эмоции



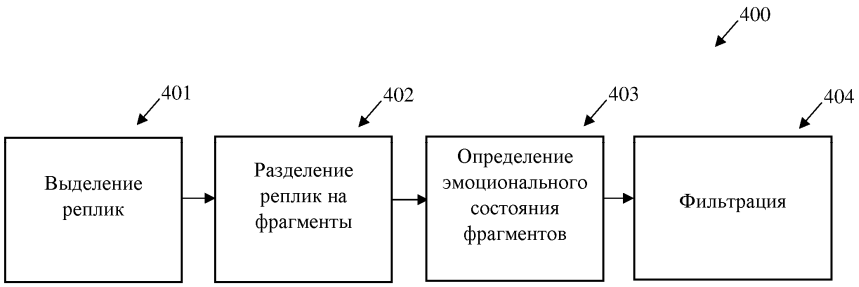
Фиг. 1



Фиг. 2



Фиг. 3



Фиг. 4

Евразийская патентная организация, ЕАПВ
Россия, 109012, Москва, Малый Черкасский пер., 2