УДК81.93.29
https://doi.org/10.47533/2020.1606-146X.09

О.*МАМЫРБАЕВ[1], А. АХМЕДИЯРОВА[1], А. КЫДЫРБЕКОВА[1,2], Н.МЕКЕБАЕВ[1,2], М. ТУРДАЛЫУЛЫ[1,2]

[1]Институт информационных и вычислительных технологий,
[2]Казахский национальный университет им. аль-Фараби

ИДЕНТИФИКАЦИЯ ГОЛОСА НА ОСНОВЕ *i*-ВЕКТОРА И ГЛУБОКИХ НЕЙРОННЫХ СЕТЕЙ С ИСПОЛЬЗОВАНИЕМ КОРОТКИХ ВЫСКАЗЫВАНИЙ

Независимое от текста распознавание пользователя по голосу с использованием коротких высказываний является очень сложной задачей из-за большого разброса и несоответствия содержания между короткими высказываниями, чтобы улучшить распознование пользователя по голосу, планируется выделить несколько наборов отличительных признаков, которые содержат больше информации, связанной с голосом. Системы, основанные на i-векторе и вероятностном линейном дискриминантном анализе (PLDA), стали стандартом в приложениях верификации пользователя по голосу, но они менее эффективны при коротких высказываниях. Результаты показывают, что система i-вектор DNN превосходит систему i-вектор GMM при различной длительности. Тем не менее, характеристики обеих систем значительно ухудшаются по мере уменьшения продолжительности высказываний. Чтобы решить эту проблему, предложена два новых метода нелинейного отображения, которые обучают модели DNN отображать i-векторы, извлеченные из коротких высказываний, в соответствующие им i-векторы длинных высказываний.

Ключевые слова: *i-вектор, глубокие нейронные сети, короткие высказывания, нелинейное отображение.*

Введение. В статье сравниваем два современных метода обучения универсальной фоновой модели (UBM) для *i*-векторного моделирования с использованием задач полной и короткой оценки высказываний. Этими двумя методами являются методы, основанные на модели *i*-вектор GMM и методы, основанные на *i*-вектор глубоких нейронных сетях (DNN). Результаты показывают, что система *i*-вектор DNN превосходит систему *i*-вектор GMM при различной длительности. Тем не менее, характеристики обеих систем значительно ухудшаются по мере уменьшения продолжительности высказываний. Чтобы решить эту проблему, мы предлагаем два новых метода нелинейного отображения, которые обучают модели DNN отображать *i*-векторы, извлеченные из коротких высказываний, в соответствующие им *i*-векторы длинных высказываний. Преобразованный *i*-вектор может восстановить недостающую информацию и уменьшить дисперсию исходных *i*-векторов с коротким высказыванием. Предложенные методы одновременно моделируют совместное представление коротких и длинных *i*-векторов высказывания: первый метод обучает автоэнкодер сначала с использованием сцепленных коротких и длинных векторов высказывания, а затем использует предварительно обученные веса для инициализации контролируемой регрессионной модели из короткого на длинную версию; второй метод совместно обучает контролируемую модель регрессии с помощью автоматического кодера, воссоздающего сам *i*-вектор короткого высказывания.

Основанная на *i* – вектор структура определила современное состояние для независимого от текста распознавания говорящего. *i*-векторы извлекаются либо из модели на основе гауссовой смеси (GMM) [1], либо из системы на основе глубокой

*Адрес для переписки. E-mail: nurbapa@mail.ru

нейронной сети (DNN) [2], а для бэкэнда – вероятностный линейный дискриминантный анализ (PLDA) используется широко. Для решения этой проблемы был изучен ряд методов по различным аспектам этой проблемы был целый ряд методов для моделирования вариации коротких *i*-векторов высказывания. В последнее время было предложено несколько подходов, в которых используются глубокие нейронные сети для обучения встраиванию пользователя по коротким высказываниям. Авторы [3] используют нейронную сеть, которая обучена различать большое количество пользователей, генерировать встроенные голосов фиксированного размера, а встроенные голоса используются для оценки PLDA. Подробно обсудим несколько ключевых факторов предлагаемых моделей отображения DNN, включая итерацию перед обучением, веса регуляризации и глубину кодера. Результаты сопоставления для систем *i*-вектор GMM и *i*-вектор DNN сравниваются и показывают значительное улучшение для обеих систем.

Системы проверки пользователей по голосу. Современная система проверки текста, независимая от текста, основана на платформе *i*-вектор. В этих системах универсальная фоновая модель (UBM) используется для сбора достаточной статистики для извлечения *i*-вектора, а вероятностный линейный дискриминантный анализ используется для получения оценок сходства между *i*-векторами. Существует два разных способа моделирования UBM: использование обученных GMM без присмотра или использование DNN, обученного в качестве классификатора senone. Поэтому мы представим системы *i* –вектор GMM и *i*-вектор DNN, а также моделирование – вероятностный линейный дискриминантный анализ.

1) *i*-вектор GMM система. Представление *i*-вектора основано на концепции моделирования полной изменчивости, которая предполагает, что зависящие от динамика и канала переменные находятся в низкоразмерном подпространстве, представленном матрицей полной изменчивости *T*. Математически супервектор GMM, *M* – множество объектов распознавания может быть смоделирован как:

$$M = m + Tw \tag{1}$$

m – независимый от голоса пользователя и сессии супервектор (который можно принять за супервектор UBM), *T* – прямоугольная матрица низкого ранга, *w* – случайный вектор, имеющий стандартное нормальное распределение $N(0; 1)$. Компоненты вектора *w* являются суммарными факторами, называем эти векторы как единичные или *i* – векторы для краткости. Чтобы узнать подпространство полной изменчивости, статистика Баума – Уэлча должна быть вычислена для данного высказывания, которые определены как:

$$N_c = \sum_t P(c|y_t, \Omega) \tag{2}$$

$$F_c = \sum_t P(c|y_t, \Omega) y_t \tag{3}$$

где N_c и F_c представляют статистику нулевого и первого порядка, y_t – выборку признаков в момент времени *t*, Ω представляют UBM компонентов смеси *C*, $c = 1, ...,$ *C* – индекс Гаусса, а $P(c|y_t, \Omega)$ соответствует апостериорной составляющей смеси *c*, порождающей вектор y_t. Чтобы оценить вектор, нам также необходимо вычислить централизованную статистику Баума-Уэлча первого порядка на основе компонентов

средней смеси UBM. Кроме того, необходимо вычислить централизованную статистику Баума – Уэлча первого порядка, основанную на математических ожиданиях универсальной фоновой модели:

$$\tilde{F}_c = \sum_{t=1}^{L} P(c|y_t, \lambda_{UBM})(y_t - \mu_c), \tag{4}$$

где μ_c – математическое ожидание компонента Гауссовой смеси *c*.

2) **i**-вектор DNN система. Для системы вектор GMM – задняя часть компонента смеси *c*, генерирующего вектор y_t, вычисляется с помощью акустической модели GMM, обученной неконтролируемым образом.

$$P(c|y_t, \Omega) \Rightarrow P(c|y_t, \Theta) \tag{5}$$

Однако в последнее время, вдохновленный успехом акустических моделей DNN в автоматическом распознавании речи (ASR), [2] предложили метод, который использует постериумы DNN сенонам (кластер контекстно-зависимых трифонов) для замены постериодов GMM, как показано в уравнении 5, что приводит к значительному улучшению верификации пользователей. Θ-представляет обученную модель DNN для классификации по сенсонам.

PLDA моделирование. PLDA – это генеративная модель распределения *i*-вектора для верификации пользователя. Используем упрощенный вариант PLDA, называемый G-PLDA, который широко используется исследователями. Стандартный G-PLDA предполагает, что *i*-вектор w_i представлен следующим образом:

$$w_i = r + U_x + \in_i \tag{6}$$

r – среднее значение *i*-векторов, *U* – определяет подпространство между голосами, *x* – латентная переменная, представляет идентичность голос пользователя и, как предполагается, имеет стандартное нормальное распределение.

$\in_i$ – остаточный член, представляет изменчивость внутри голоса, которая обычно распределяется с нулевым средним и полной ковариацией Σ' . Оценка системы *i*-вектора на основе PLDA рассчитывается с использованием логарифмического отношения правдоподобия (LLR) между целевым и тестовым *i*-векторами, которое обозначается как $w_{\text{цель}}$ и $w_{\text{тест}}$. Отношение правдоподобия можно рассчитать следующим образом:

$$LLR = \log \frac{P(w_{\text{цель}}, w_{\text{тест}}|H_1)}{P(w_{\text{цель}}|H_0)}, \tag{7}$$

где H_1 и H_0 обозначают гипотеза о том, что два *i*-вектора представляют одного и того же говорящего и разных говорящих соответственно. Чтобы показать изменения вариаций между длинным и коротким *i*-векторами высказывания, сначала вычисляем среднюю диагональную ковариацию (обозначенную как σ_m) *i*-векторов по всем высказываниям данного говорящего *m*, а затем вычисляем среднее значение (обозначаемое как означает) ковариаций по всем пользователям. σ_m и σ_{mean} определены в формулах 8 и 9 как:

$$\sigma_m = \frac{1}{N} \sum_{n=1}^{N} T_r((w_{mn} - \bar{w}_m)(w_{mn} - \bar{w}_m)^T) \tag{8}$$

$$\sigma_{mean} = \frac{1}{M}\sum_{m=1}^{M}\sigma_m , \tag{9}$$

где $\overline{w}_m$ соответствует среднему значению *i*-векторов, принадлежащих пользователю *m*. *N* представляет общее количество высказываний для пользователя *m*, T_r представляет операцию трассировки, а *M* – общее количество голосов. Чтобы сравнить среднее значение для длинных и коротких *i*-векторов высказывания, выбираем около 200 голос пользователей с множеством длинных высказываний.

i-векторное отображение основано на DNN. Чтобы уменьшить возможное несоответствие фонем в коротких текстозависимых высказываниях, предлагаем несколько способов сопоставить векторы коротких высказываний с их длинной версией. Эксперименты показывают, что контролируемый DNN начинает непосредственно изучать это отображение, что аналогично подходам в [4], улучшение не является значительным из-за чрезмерной подгонки к обучающему набору данных. Первый – это двухэтапный метод: с помощью авто-кодера сначала обучают представлению узкого места как длинных, так и коротких *i*-векторов высказывания, а затем используют предварительно обученные веса для выполнения контролируемой тонкой настройки модели, которая преобразует *i*-вектор короткого высказывания в его длинную версию напрямую. Второй – это одностадийный метод: совместно обучить контролируемую модель регрессии с помощью автоматического кодера для восстановления короткого *i*-вектора. Конечная потеря для оптимизации представляет собой взвешенную сумму контролируемой потери регрессии и потери реконструкции. Далее подробно рассмотрим эти два метода.

DNN_1 (двухэтапный метод): предварительная подготовка и точная настройка. Чтобы найти хорошую инициализацию модели DNN под наблюдением, сначала обучаем совместное представление как коротких, так и длинных векторов высказывания, используя автоэнкодер. Сначала объединяем короткий *i*-вектор w_s и его длинную версию w_i в *z*, затем каскадный вектор *z* используется для обучения автоэнкодера с некоторыми конкретными ограничениями. Автоэнкодер состоит из кодера и декодера. Функция кодера $h = f(z)$ изучает скрытое представление входного вектора *z*, а функция декодера $\hat{z} = g(h)$ производит реконструкцию. Процесс обучения описывается как минимизация функции потерь $L(z, g(f(z)))$. Чтобы узнать более полезное представление, добавляем ограничение на автоэнкодер: ограничьте скрытое представление *h* относительно небольшим измерением, чтобы изучить наиболее характерные особенности обучающих данных. Для функции кодера $f(z)$ принимаем варианты из нескольких, полностью связанных слоев для сложенных остаточных блоков [5], чтобы исследовать влияние глубины кодера. Для функции декодера $g(h)$ используем один, полностью связанный слой с слоем линейной регрессии, так как этого достаточно, чтобы приблизить отображение из изученного скрытого представления *h* к выходному вектору. Для функции потерь используем критерий среднеквадратичной ошибки, который равен $\left\|g(f(z)) - z\right\|^2$. После того, как автоэнкодер обучен, используем обученную DNN-структуру и веса для инициализации контролируемого отображения. Оптимизируем потери между предсказанным длинным вектором и действительным длинным i-вектором

$$\frac{1}{N}\sum_{n=1}^{N}\left\|\hat{w}_l - w_l\right\|^2 \tag{10}$$

Обозначим этот метод как DNN_1.

DNN_2 (одноступенчатый метод): обучение с полудонтролем. Двухэтапный метод, упомянутый в предыдущем разделе, должен сначала обучить совместному представлению с использованием автоэнкодера, а затем выполнить точную настройку для обучения контролируемого отображения. Вводим еще одну унифицированную среду с полууправлением, основанную на нашей предыдущей работе [6], которая может совместно обучать контролируемое отображение с помощью автоматического кодера для минимизации ошибки восстановления. Этот метод обозначается как DNN_2. Принимаем ту же платформу автоэнкодера, как упомянуто в предыдущем разделе, в которой есть кодер и декодер, но входом для кодера здесь является *i*-вектор с коротким произнесением w_s. Выходной сигнал от кодера будет подключен к слою линейной регрессии для прогнозирования *i*-вектора длинных высказываний w_l, и он также будет использоваться для восстановления самого *i*-вектора коротких высказываний путем ввода его в декодер, что дает подняться к структуре автоэнкодера. Определяем новую целевую функцию для совместного обучения сети. Используем $\hat{w}_l$ и $\hat{w}_s$ для представления выходных данных модели контролируемой регрессии и автоэнкодера соответственно. Можем определить функцию объективных потерь $L_{общ}$, которая объединяет потери из регрессионной модели и автоэнкодера взвешенным способом как:

$$L_{общ} = (1 - \alpha)L_r + \alpha L_a \tag{11}$$

где L_r – модель потери регрессии, определяемая как

$$L_r(w_s, w_l; \theta_r) = \frac{1}{N}\sum_{n=1}^{N}\left\|\hat{w}_l - w_l\right\|^2 \tag{12}$$

и L_a – потеря автокодера, определяемая как:

$$L_a(w_s, w_s; \theta_a) = \frac{1}{N}\sum_{n=1}^{N}\left\|\hat{w}_s - w_s\right\|^2 \tag{13}$$

Кроме того, θ_r и θ_a являются параметрами регрессионной модели и автоматического кодера соответственно, которые совместно обучаются и совместно используют веса уровня кодера. *a* - скалярный вес, который определяет, насколько ошибка реконструкции используется для регуляризации контролируемого обучения. Потеря реконструкции автоэнкодера L_a заставляет скрытый вектор, сгенерированный из кодера, реконструировать *i*-вектор короткого высказывания в дополнение к прогнозированию целевого *i*-вектора длинного высказывания w_l, и помогает предотвратить переопределение скрытого вектора w_l. Для тестирования используем только выход из регрессионной модели w_l в качестве отображенного *i*-вектора.

Процедура обучения по-прежнему соответствует предлагаемым методам совместного моделирования (DNN_1 или DNN_2). Ожидается, что векторы фонем помогут нормализовать *i*-вектор коротких высказываний и обеспечат дополнительную информацию для этого отображения. Фонемный вектор *p* определяется как:

$$p = \frac{1}{N}\sum_{t=1}^{N} P(c|y_t, \Theta) \tag{14}$$

Задний $P(c|y_t, \Theta)$ генерируется из основанного на TDNN классификатора сенона, который был определен выше.

Экспериментальная установка. *i*-векторные базовые системы. Оцениваем наши методы с использованием самых современных систем *i*-вектор/G-PLDA на основе GMM и DNN с использованием инструментария Kaldi [7]. Обучение *i*-векторному картированию. Все нейронные сети обучаются с использованием стратегии оптимизации Адама [8] с критерием среднеквадратичной ошибки и экспоненциально убывающей скоростью обучения, начиная с 0,001.

i-векторные базовые системы. Представляем и сравниваем две базовые системы: систему -вектор GMM и систему *i*-вектор DNN со стандартным полноразмерным условием NIST SRE 10 и усеченными условиями 10c-10c и 5c-5c. Когда обе системы оценивались в усеченных условиях оценки 10c-10c, 5c-5c, характеристики значительно ухудшаются по сравнению с условиями полной длины. Основная причина в том, что когда длина оценочного высказывания короче, между высказываниями возникает значительное фонетическое несоответствие. Однако производительность системы *i*-вектор DNN по-прежнему превосходит систему *i*-вектор GMM на 8–24%, хотя улучшение не такое большое, как в условиях полной длины. Влияние потерь при восстановлении для DNN_2 – исследуем влияние весов на потери при восстановлении в DNN_2. Следовательно, кажется, что неконтролируемое обучение очень важно для задачи распознавания голоса пользователя. Поэтому сравним неглубокий (2-слойный) и глубокий (5-слойный) кодеры для DNN_1 и DNN_2 и используем два метода для решения этой проблемы. Во-первых, используем нормализованную инициализацию и пакетный уровень нормализации для нормализации промежуточного скрытого вывода. Во-вторых, применяем остаточное обучение, которое использует несколько остаточных блоков без дополнительных параметров по сравнению с обычными полностью связанными слоями. Эффект от добавления информации о фонемах – показываем результаты при добавлении вектора фонем (среднее значение постер фонем через кадры) с кратким произнесением *i*-векторов, чтобы выучить отображение. Исследуем эффект добавления информации о фонем на основе наилучших структур отображения DNN. Результаты показывают, что предлагаемые методы могут быть использованы в реальных условиях, таких как умный дом и криминалистические приложения.

Эффекты картирования. Чтобы исследовать влияние предложенных алгоритмов отображения *i*-вектора, сначала рассчитываем среднее квадратное евклидово расстояние между короткими и длинными *i*-векторными парами высказываний в наборе данных оценки SRE10 до и после отображения. Среднеквадратичное евклидово расстояние D_{sl} между коротким и длинным i-вектором произнесения определяется следующим образом:

$$D_{sl} = \frac{1}{N}\sum_{s=1}^{N}\left(\sum_{i=1}^{L}(w_s(i) - w_i(i))^2\right), \tag{15}$$

где w_s и w_i представляют *i*-вектор короткого и длинного высказывания соответственно, L – длина *i*-вектора, а N – число коротких и длинных пар *i*-вектора. Сравниваем значения D_{sl} для 10-секундных и 5-секундных *i*-векторов короткого высказывания, а также сопоставленные 10-секундные и 5-секундные *i*-векторы короткого высказывания для женщин и мужчин. Отображенные *i*-векторы с коротким высказыванием имеют значительно меньший D_{sl} по сравнению с теми, что были до отображения. После отображения D_{sl} в состоянии 10 с меньше, чем в состоянии 5 с. Кроме того, вычисляем и сравниваем *j*-коэффициент [9] *i*-векторов с коротким произнесением в Таблице 1,

которая измеряет способность разделения голосов. Учитывая *i*-векторы для *M* голосов, *j*-коэффициент может быть вычислен с использованием уравнений. (16) - (18):

$$S_w = \frac{1}{M}\sum_{s=1}^{M} R_i \tag{16}$$

$$S_w = \frac{1}{M}\sum_{s=1}^{M} (w_i - w_0)(w_i - w_0)^T \tag{17}$$

$$j = T_r((S_b + S_w)^{-1} S_b), \tag{18}$$

где S_w – матрица рассеяния внутри класса, S_b – матрица рассеяния между классами, w_i – средний *i*-вектор означает среднее значение всех значений w_i, R_i – ковариационную матрицу для *i*-го голоса.

Из Таблицы 1 можем наблюдать, что отображенные *i*-векторы имеют значительно более высокие *j*-отношения по сравнению с оригинальными векторами с коротким произнесением для условий 5с и 10с. Эти результаты указывают на то, что предлагаемые методы отображения на основе DNN могут хорошо обобщаться на невидимые голоса и высказывания и улучшать способность *i* -вектора к разделению голоса.

Таблица 1 – *j*-отношения по сравнению с оригинальными векторами

	10с		5с	
	оригинал	картирования	оригинал	картирования
женщина	90.96	93.56	88.05	89.76
мужчина	87.45	90.07	86.09	87.87

Выводы. Быстродействие систем верификации голоса *i*-вектор на основе GMM и DNN быстро ухудшается по мере уменьшения продолжительности оценочных высказываний. Это объясняет и анализирует причины ухудшения и предлагает несколько методов, основанных на DNN, для обучения нелинейного отображения *i*-векторов с коротким высказыванием на их длинную версию, чтобы улучшить производительность оценки короткого высказывания.

Предложены два метода отображения на основе DNN (DNN_1 и DNN_2), и они оба моделируют совместные представления *i*-векторов с коротким или длинным высказыванием. Как DNN_1, так и DNN_2 приводят к значительному улучшению по сравнению с базовой линией оценки коротких высказываний для голоса мужчин и женщин, а также с большим отрывом превосходят другие методы компенсации коротких высказываний. После проведения *t*-теста ($p < 0{,}001$) результаты показывают, что все улучшения являются статистически значимыми. Изучаем несколько ключевых факторов моделей DNN и заключаем следующее:

1. Для обученной модели DNN с полууправлением (DNN_2) неконтролируемое обучение играет более важную роль, чем контролируемое обучение в задаче проверки голос пользователя;

2. Увеличивая глубину нейронных сетей, используя остаточные блоки, можем облегчить проблему жесткой оптимизации глубоких нейронных сетей и получить улучшение по сравнению с мелкой сетью, особенно для DNN_1;

3. Предложенные методы отображения на основе DNN хорошо работают для коротких высказываний с различной и смешанной длительностью;

4. Предложенные модели могут также улучшить системы *i*-вектор GMM и *i*-вектор DNN, и после отображения система *i*-вектор DNN по-прежнему работает лучше, чем система *i*-вектор GMM и дают значительное улучшение для коротких высказываний произвольной длины.

Данная работа была поддержана Министерством образования и науки Республики Казахстан. IRN AP05131207 Разработка технологий многоязычного автоматического распознавания речи с использованием глубоких нейронных сетей.

ЛИТЕРАТУРА

1 Dehak, N., Кенни, P.J., Dehak, R., Dumouchel, P., Ouellet, P., 2011. Фронтальный факторный анализ для проверки докладчика. IEEE Trans. Аудио Речь Ланг. Процесс. – № 19 (4), с. 788–798. [Dehak, N., Kenni, P.J., Dehak, R., Dumouchel, P., Ouellet, P., 2011. Frontal'nyj faktornyj analiz dlya proverki dokladchika. IEEE Trans. Audio Rech' Lang. Process. – № 19 (4), s. 788–798.]

2 Li, L., Wang, D., Zhang, C., Zheng, T.Z. , 2016. Улучшение распознавания коротких высказываний за счет моделирования классов речевых единиц. IEEE Trans. Аудио Речь Ланг. Процесс. – № 24 (6), с. 1129–1139. [Li, L., Wang, D., Zhang, C., Zheng, T.Z. , 2016. Uluchshenie raspoznavaniya korotkih vyskazyvanij za schet modelirovaniya klassov rechevyh edinic. IEEE Trans. Audio Rech' Lang. Process. – № 24 (6), s. 1129–1139.]

3 Снайдер Д., Гарсия-Ромеро Д., Повей Д., 2015. Временная задержка универсальных фоновых моделей на основе нейронной сети для распознавания говорящего. Материалы семинара IEEE по автоматическому распознаванию и пониманию речи, с. 92–97.[Snajder D., Garsiya-Romero D., Povej D., 2015. Vremennaya zaderzhka universal'nyh fonovyh modelej na osnove nejronnoj seti dlya raspoznavaniya govoryashchego. Materialy seminara IEEE po avtomaticheskomu raspoznavaniyu i ponimaniyu rechi, s. 92–97.]

4 Bousquet, P.M., Rouvier, M., 2017. Компенсация несоответствия по длительности с использованием четырехковариантной модели и глубокой нейронной сети для проверки диктора. Материалы Interspeech, с. 1547–1551. [Bousquet, P.M., Rouvier, M., 2017. Kompensaciya nesootvetstviya po dlitel'nosti s ispol'zovaniem chetyrekhkovariantnoj modeli i glubokoj nejronnoj seti dlya proverki diktora. Materialy Interspeech, s. 1547–1551.]

5 He, K., Zhang, X., Ren, S., Sun, J. 2016. Deep residual learning for imagerecognition, in Proc. of the IEEE conference on computer vision and patternrecognition 2016. – P. 770 –778.

6 О.Ж.Мамырбаев, М.Отман, А.Т.Ахмедиярова, А.С.Кыдырбекова, Н.О.Мекебаев «Голосовая верификация с использованием i-векторов и нейронных сетей с ограниченными данными обучения» Бюллетень Национальной академии наук РК Выпуск: 3, 2019, с.36-43. [O.ZH.Mamyrbaev, M.Otman, A.T.Ahmediyarova, A.S.Kydyrbekova, N.O.Mekebaev «Golosovaya verifikaciya s ispol'zovaniem i-vektorov i nejronnyh setej s ogranichennymi dannymi obucheniya» Byulleten' Nacional'noj akademii nauk RK Vypusk: 3, 2019, s.36-43.]

7 Повей Д., Гошаль А., Булянн Г., Бургет Л., Глембек О., Гоэль Н., Силовский Дж., 2011. Инструментарий распознавания речи Kaldi. Материалы семинара по распознаванию и пониманию речи IEEE. Prince, S.J., Elder, J.H., 2007. Вероятностный линейный дискриминантный анализ для выводов об идентичности. Слушания ICCV, 1–8.[Povej D., Goshal' A., Bulyann G., Burget L., Glembek O., Goel' N., Silovskij Dzh., 2011. Instrumentarij raspoznavaniya rechi Kaldi. Materialy seminara po raspoznavaniyu i ponimaniyu rechi IEEE. Prince, S.J., Elder, J.H., 2007. Veroyatnostnyj linejnyj diskriminantnyj analiz dlya vyvodov ob identichnosti. Slushaniya ICCV, 1–8.]

8 Кингма Д., Ба, Дж., 2014. Адам: метод стохастической оптимизации. arXiv: 1412.6980 Lei, Y., Scheffer, N., Ferrer, L., McLaren, M., 2014. Новая схема распознавания говорящего с использованием фонетически осведомленной глубокой нейронной сети. Слушания ICASSP, 1695–1699. [Kingma D., Ba, Dzh., 2014. Adam: metod stohasticheskoj optimizacii. arXiv: 1412.6980 Lei, Y., Scheffer, N., Ferrer,

L., McLaren, M., 2014. Novaya skhema raspoznavaniya govoryashchego s ispol'zovaniem foneticheski osvedomlennoj glubokoj nejronnoj seti. Slushaniya ICASSP, 1695–1699.]

9 Фукунага, К., 1990. Введение в статистическое распознавание образов. Академическая пресса. Glorot, X., Bengio, Y., 2010. Понимание сложности обучения глубоких нейронных сетей с прямой связью. Труды тринадцатой Международной конференции по искусственному интеллекту и статистике. – С. 249–256.[Fukunaga, K., 1990. Vvedenie v statisticheskoe raspoznavanie obrazov. Akademicheskaya pressa. Glorot, X., Bengio, Y., 2010. Ponimanie slozhnosti obucheniya glubokih nejronnyh setej s pryamoj svyaz'yu. Trudy trinadcatoj Mezhdunarodnoj konferencii po iskusstvennomu intellektu i statistike. – S. 249–256.]

О. МАМЫРБАЕВ[1], А. АХМЕДИЯРОВА[1], А. ҚЫДЫРБЕКОВА[1,2], Н. МЕКЕБАЕВ[1, 2], М. ТҰРДАЛЫ[1, 2]

[1]Ақпараттық және есептеуіш технологиялар институты,
[2]Әл-Фараби атындағы қазақ ұлттық университеті

ҚЫСҚА МӘЛІМДЕМЕЛЕРДІ ҚОЛДАНА ОТЫРЫП, І-ВЕКТОРҒА ЖӘНЕ ТЕРЕҢ НЕЙРОНДЫҚ ЖЕЛІЛЕРГЕ НЕГІЗДЕЛГЕН ДАУЫСТЫ АНЫҚТАУ

Қысқа сөйлемдерді пайдалана отырып дауыс арқылы тану – қысқа сөйлемдер арасындағы мазмұнның кең таралуы мен сәйкес келмеуіне байланысты өте қиын міндет, өйткені пайдаланушыны дауыс арқылы тануды жақсарту үшін, дауысқа қатысты көбірек ақпаратты қамтитын бірнеше ерекшеліктерді бөліп көрсету жоспарланған. Жүргізілген зерттеу нәтижелері і-векторлы DNN жүйесінің әртүрлі ұзындықтарға арналған GMM і-векторлық жүйеден жоғары екенін көрсетеді. Алайда, екі жүйенің сипаттамалары сөйлемдердің ұзақтығы қысқарған сайын айтарлықтай нашарлайды. Бұл мәселені шешу үшін DNN модельдерін қысқа сөйлемдерден алынған ұзын сөйлемдердің сәйкес келетін і-векторларына салыстыруға үйрететін екі жаңа сызықты емес әдісті ұсынамыз.

Түйін сөздер**:** *і-вектор, терең нейрондық желілер, қысқа сөйлемдер, сызықты емес әдіс.*

O. MAMYRBAYEV[1], A. AKHMEDIYAROV[1], A. KYDYRBEKOV[1, 2], N. MEKEBAYEV[1, 2], M. TURDALYULY[1, 2]

[1]Institute of information and computing technologies,
[2]Kazakh al-Farabi national university

VOICE IDENTIFICATION BASED ON THE I-VECTOR AND DEEP NEURAL NETWORKS USING SHORT UTTERANCES

Text-independent voice recognition of the user using short sentences is a very difficult task due to the large spread and inconsistency of the content between short sentences, in order to improve user recognition by voice, it is planned to highlight several sets of distinguishing features that contain more information related to the voice. The results show that the i-vector DNN system is superior to the GMM i-vector system for various durations. However, the characteristics of both systems deteriorate significantly as the duration of the sentences decreases. To solve this problem, we propose two new nonlinear mapping methods that train DNN models to map i-vectors extracted from short sentences to their corresponding i-vectors of long sentences.

Key words: *i-vector, deep neural networks, short utterances, nonlinear mapping.*