УДК 004.934

О.И. Федяев, И.Ю. Бондаренко

Нейросетевой алгоритм дикторонезависимого распознавания фонем устной речи

Рассмотрена проблема дикторонезависимого распознавания фонем устной речи, возникающая при создании систем автоматического распознавания слов дискретной и слитной речи на основе фонемно-ориентированного метода. Проведен сравнительный анализ результатов данных экспериментов с результатами, полученными при использовании скрытых марковских моделей.

This paper explains the speaker independent recognition of oral speech phonemes problem which appears in creation of automated recognition systems for words of discrete and continuous speech using the phoneme-oriented method. The experiments' data was comparatively analyzed with results taken using hidden Markov models.

Розглянуто проблему дикторонезалежного розпізнання фонем усного мовлення, яка виникає зі створенням систем автоматичного розпізнання слів дискретного та злитого мовлення на основі фонемно-орієнтованого методу. Проведено порівняльний аналіз результатів експериментів з результатами, отриманими з використанням прихованих марківських моделей.

Введение. Типовая система распознавания слитной речи имеет структуру, состоящую из двух последовательных блоков: акустического и лингвистического [1]. Акустический блок выполняет предварительный анализ речевого сигнала, выделение признаков и распознавания структурных элементов речи (аллофонов, фонем, слогов или слов). Лингвистический – осуществляет интерпретацию акустической информации с учетом модели словаря и языка, формирует окончательный результат распознавания.

Ключевым компонентом акустического блока любой системы распознавания устной речи служит алгоритм распознавания последовательности базовых структурных элементов устной речи. В качестве таких структурных элементов наиболее часто используют фонемы, поскольку:

- фонему можно рассматривать как минимальную линейную единицу устной речи [2];
- количество фонем в каждом языке ограничено, что упрощает задачу их распознавания в речевом сигнале и процесс обучения такому распознаванию.

Существует много алгоритмов распознавания фонем в слитной речи, но все они могут быть отнесены к одному из двух классов: *генеративных* и *дискриминативных* алгоритмов распознавания.

Среди класса генеративных алгоритмов распознавания наиболее популярны скрытые Марковские модели [3] и КДП-подход [4]. Известны более-менее успешные попытки использования этих алгоритмов для пофонемного распознавания устной речи. Среди таких попыток следует отметить эксперименты по построению систем распознавания устной речи со сверхбольшим словарем для английского [5], японского [6], русского [7] и украинского [8] языков. В зависимости от языка и речевого корпуса, на котором проводилось обучение и тестирование, точность распознавания фонем в этих экспериментах составила 60–70 процентов.

Главный принцип действия как скрытых Марковских моделей, так и КДП-подхода – генерация максимально правдоподобных эталонных сигналов на основе некоторой автоматной грамматики и сопоставление полученных эталонов с распознаваемым речевым сигналом. Такой принцип обуславливает как преимущества, так и недостатки этих алгоритмов. К существенному преимуществу генеративных алгоритмов следует отнести эффективное моделирование процессов, нелинейно изменяющихся во времени, а среди недостатков можно отметить не очень высокую дискриминатную способность.

Ключевые слова. Автоматическое распознавание речи, контекстно-независимое распознавание фонем, искусственные нейронные сети.

К противоположному классу – дискриминативных алгоритмов – относятся алгоритмы, основанные на построении границ между распознаваемыми классами в пространстве признаков. Наиболее распространенным математическим аппаратом для разработки дискриминативных алгоритмов распознавания считаются искусственные нейронные сети. Основные преимущества этого математического аппарата таковы:

- многослойные нейронные сети имеют высокую дискриминантную способность;
- нейронная сеть во время обучения может найти оптимальную комбинацию ограничений для классификации образов, и при этом нет необходимости в жестких предположениях о распределении входных признаков (что необходимо, например, в скрытых Марковских моделях);
- нейросетевой алгоритм характерен хорошими скоростными данными при высокой степени параллелизма.

К недостаткам нейронных сетей можно отнести то, что с помощью этого математического аппарата трудно моделировать высокую вариативность распознаваемых сигналов во времени.

Существует ряд систем распознавания, в которых алгоритмы функционирования акустического блока основаны на нейронных сетях. Среди них следует отметить системы распознавания английского [9] и русского [10] языков, которые показали неплохие результаты распознавания фонем в диапазоне 65–70 процентов.

Таким образом, результаты исследований показывают, что, несмотря на ряд разработанных алгоритмов распознавания фонем в устной речи, проблема повышения точности такого распознавания остается актуальной.

Цель работы состоит в повышении точности распознавания структурных элементов (фонем) речевого сигнала, выполняемого акустическим блоком.

Объект исследования – акустический блок системы распознавания, а *предмет* – алгоритм автоматического дикторонезависимого распознавания фонем в слитной речи на базе математического аппарата искусственных нейронных сетей.

Нейросетевой алгоритм распознавания фонем. Это вычислительная процедура, основная часть которой может быть реализована на искусственной нейронной сети. Для построения алгоритма необходимо определить [11]:

- объект в роли входного сигнала нейронной сети;
- объект в роли выходного сигнала нейронной сети;
- желаемый выходной сигнал нейронной сети;
- структуру нейронной сети;
- функцию ошибки нейронной сети;
- критерий качества нейронной сети и функционал ее оптимизации, зависящий от ошибки;
- значения весовых коэффициентов.

Входной сигнал нейронной сети. Этот сигнал формируется следующим образом. Текущий фрагмент речевого сигнала продолжительностью T_F, который необходимо распознать (отнести к одной из фонем словаря), пропускается через цифровой фильтр с характеристикой $1 - 0{,}97z^{-1}$ во избежание влияния формы сигнала голосовой щели и характеристик излучения губ диктора [12]. Затем профильтрованный фрагмент речевого сигнала подвергается скользящему оконному анализу окном Хэмминга [13] длиной $T_{\text{window}} = 20$ мс и шагом скольжения $T_{\text{shift}} = 10$ мс. Таким образом, речевой фрагмент разбивается на окна – короткие блоки звуковых данных, расположенные последовательно вдоль оси времени. На основе данных этих блоков вычисляются векторы мел-частотных кепстральных коэффициентов (*Mel Frequency Cepstral Coefficients – MFCC*) [14]. В данной статье использованы традиционные 13-мерные *MFCC*-векторы (логарифм энергии и 12 коэффициентов *MFCC*) C_i, $i = 1\ldots13$.

Пример описания речевого сигнала, полученного в результате произнесении слова *She* (Она), с помощью логарифма энергии и мел-частотных кепстральных коэффициентов приведен на рис. 1. Транскрипция слова *She* представляет собой последовательность английских фонем *sh* и *iy*. Как видно из рисунка, на границе этих фонем существенно меняется спектральная картина речевого сигнала, описываемая мел-час-

тотными кепстральными коэффициентами. Кроме того, на участке речевого сигнала, соответствующем фонеме *iy*, резко возрастает энергия сигнала.

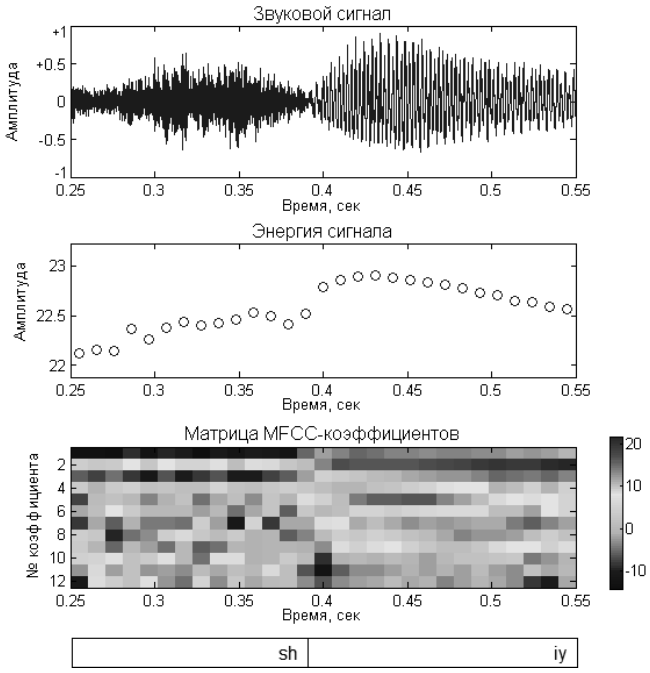

Рис. 1. Пример описания речевого сигнала, полученного при произнесении слова *She* и образующего последовательность фонем *sh* и *iy*, с помощью логарифма энергии и мел-частотных кепстральных коэффициентов

Компоненты каждого *MFCC*-вектора нормализуются так, чтобы математическое ожидание по каждому компоненту стало нулевым, а среднеквадратичное отклонение – единичным:

$$\tilde{C}_{it} = \frac{C_{it} - M_i}{\sigma_i}, \; i = 1\ldots 13, \; t = 1\ldots N_F, \qquad (1)$$

где i – номер компонента *MFCC*-вектора; t – номер окна анализа и, соответственно, *MFCC*-вектора, вычисленного в этом окне; $N_F = \text{floor}\left(\left(T_F - T_{\text{window}}\right)/T_{\text{shift}}\right)+1$ – количество окон, которое удается вычислить в речевом фрейме длиной T_F с заданными параметрами оконного анализа T_{window} и T_{shift}; M_i и σ_i – математическое ожидание и среднеквадратичное отклонение для i-го компонента *MFCC*-вектора, вычисленные на всей совокупности речевых сигналов обучающего множества. Такая нормализация необходима для того, чтобы выровнять амплитуду изменений *MFCC*-коэффициентов верхнего и нижнего порядков, соответствующих верхним и нижним частотам спектра (особенность спектра речевого сигнала – то, что амплитуда изменений энергии на верхних частотах значительно меньше, чем на нижних).

Последовательность нормализованных *MFCC*-векторов, формируемая описанным способом, представляет собой нормализованный мел-частотный кепстральный образ текущего фрагмента речевого сигнала и есть входным сигналом нейронной сети, распознающей фонемы.

Выходной сигнал нейронной сети. В качестве выходного сигнала нейронной сети выступает вектор степеней принадлежности текущего речевого фрагмента фонемам словаря Out_j, $j = 1\ldots N_{\text{phones}}$. Размер этого вектора равен количеству распознаваемых фонем N_{phones}. Каждый компонент вектора соответствует «своей» фонеме словаря. Результат распознавания – номер фонемы в словаре, степень принадлежности речевого фрагмента к которой максимальна – определяется следующим образом:

$$j_{\text{res}} = \arg\max_{j=1:N_{\text{phones}}} \left(\text{Out}_j\right). \qquad (2)$$

Желаемый выходной сигнал нейронной сети. Желаемый сигнал представляет собой вектор размером N_{phones}, формируемый следующим образом:

- компонент вектора, соответствующий фонеме, которая в наибольшей степени представлена в текущем речевом фрагменте и, соответственно, должна быть распознана, принимает значение +1;
- остальные компоненты вектора принимают значение –1.

Процедура определения фонемы, в наибольшей степени представленной в текущем речевом фрагменте, организована так, как описано ниже.

Пусть $(t_{\text{start}}; t_{\text{start}} + T_F)$ – границы текущего фрагмента речевого сигнала, $PHONES = \{\text{Phones}_i\}$, $i = 1\ldots N_{\text{trans}}$ – транскрипция речевого сигнала, состоящая из N_{trans} фонем, $BORDERS = \left\{\left(T_i^{\text{start}}; T_i^{\text{end}}\right)\right\}$, $i = 1\ldots N_{\text{trans}}$ – множество временных ме-

ток, задающих границы этих фонем в речевом сигнале, а $DIST = \{\text{dist}_i\}$, $i = 1 \ldots N_{\text{trans}}$ – расстояния от центров этих фонем до центра текущего фрагмента речевого сигнала, вычисляемые по следующей формуле:

$$\text{dist}_i = \left| t_{\text{start}} + \frac{T_F - \left(T_i^{\text{start}} + T_i^{\text{end}}\right)}{2} \right|, \quad i = 1 \ldots N_{\text{trans}} \quad (3)$$

Фонема $\text{Phones}_{i_{\text{trg}}}$ считается фонемой в наибольшей степени представленной в текущем речевом фрагменте $(t_{\text{start}}; t_{\text{start}} + T_F)$, тогда и только тогда, когда центр этой фонемы расположен ближе к центру текущего речевого фрагмента, чем центры других фонем в транскрипции:

$$i\_\text{trg} = \arg\max_{i=1:N_{\text{trans}}} \left(\text{dist}_i\right). \quad (4)$$

На рис. 2. показано, как происходит определение фонемы, в наибольшей степени представленной в текущем речевом фрагменте. Здесь изображены:

- звуковой сигнал, полученный в результате произнесения английского слова *She* (Она) с паузой перед словом;
- фонетическая транскрипция этого сигнала $PHONES = \{h\#,\ sh,\ iy\}$;
- множество временных меток, задающих границы фонем в сигнале, $BORDERS = \{(0{,}00\text{ с}; 0{,}14\text{ с}); (0{,}14\text{ с}; 0{,}39\text{ с}); (0{,}39\text{ с}; 0{,}57\text{ с})\}$.

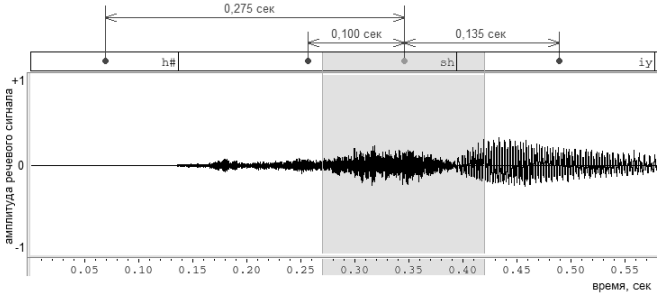

Рис. 2. Определение фонемы, в наибольшей степени представленной в текущем речевом фрагменте звукового сигнала (текущий фрагмент выделен серым цветом)

Текущий фрагмент речевого сигнала (0,27 с; 0,42 с) показан на рисунке серым цветом (рис. 2). Согласно формуле (3) вычисляем расстояния от центров всех фонем транскрипции до центра текущего фрагмента речевого сигнала $DIST = \{0{,}275\text{ с}; 0{,}100\text{ с}; 0{,}135\text{ с}\}$. Затем, применяя формулу (4), получаем, что $i\_\text{trg} = 2$. Таким образом, несмотря на то, что текущий фрагмент речевого сигнала захватывает как фонему *sh*, так и фонему *iy*, в наибольшей степени в этом фрагменте представлена все-таки фонема *sh* – вторая фонема транскрипции. Соответственно, при подготовке обучающего множества в данном примере вектор желаемого выходного сигнала нейронной сети будет формироваться так, что компонент этого вектора, соответствующий фонеме *sh*, примет значение $+1$, а другие компоненты, соответствующие остальным фонемам словаря, примут значение -1.

Структура нейронной сети. Для распознавания фонем была применена нейронная сеть типа многослойный персептрон, имеющая классическую многослойную структуру с полными последовательными связями и сигмоидальными функциями активации нейронов. Известно, что двухслойный персептрон способен аппроксимировать сколь угодно сложную непрерывную функцию, в том числе и функцию, которая описывает нелинейную гиперповерхность, разделяющую в пространстве признаков распознаваемые классы. Однако более эффективным аппроксиматором будет трехслойный персептрон, особенно если распознаваемые классы образуют в пространстве признаков сложные многосвязные области [15]. Исходя из этого, для распознавания фонем принято решение использовать многослойный персептрон с тремя слоями нейронов – двумя скрытыми и выходным (рис. 3). Размер первого скрытого слоя N_1 и размер второго скрытого слоя N_2 определяются экспериментально, а размер выходного слоя всегда равен числу распознаваемых классов фонем N_{phones}.

Для реализации сигмоидальной функции активации нейронов использована рациональная сигмоида следующего вида:

$$f(x) = \frac{2 \cdot x}{1 + |x|}. \quad (5)$$

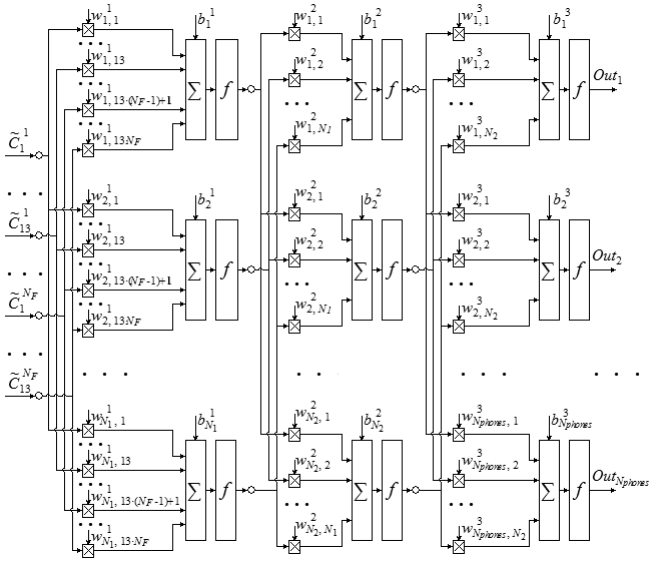

Рис. 3. Структурная схема нейронной сети, решающей задачу распознавания фонем

Такая функция активации имеет следующие преимущества:

- в отличие от логистической сигмоиды, данная функция – биполярна, и, соответственно, все сигналы в нейронной сети также будут биполярными, что позволяет уменьшить число итераций алгоритма обучения этой сети [16];
- в сравнении с другой биполярной сигмоидой – гиперболическим тангенсом – данная функция вследствие своей простоты гораздо быстрее вычисляется на ЭВМ, что позволяет ускорить функционирование нейронной сети как в режиме распознавания, так и в режиме обучения.

Математическая модель нейронной сети, используемой в данной статье, описывается следующим выражением:

$$Out_j = f\left(b_{j3} + \sum_{k_3=1}^{N_2} w^3_{j,k_3} f\left(b^2_{k_3} + \sum_{k_2=1}^{N_1} w^2_{k_3,k_2} f\left(b^1_{k_2} + \sum_{k_1=1}^{13\cdot N_F} w^1_{k_2,k_1} \tilde{C}^t_i\right)\right)\right), \quad (6)$$

где $j = 1 \ldots N_{\text{phones}}$ – индекс компонента вектора выходного сигнала нейронной сети, а $t = \text{floor}(k_1/13) + 1$, $i = k_1 - 13 \cdot \text{floor}(k_1/13)$ – индексы компонента нормализованного мел-частотного кепстрального образа, служащего входным сигналом нейронной сети (здесь floor(·) – оператор, возвращающий ближайшее меньшее целое от числа).

Функция ошибки и функционал обучения нейронной сети. Данная функция определяется как сумма квадратов расстояний от компонентов выходного сигнала до соответствующих им компонентов желаемого выходного сигнала:

$$E = 0,5 \cdot \sum_{j=1}^{N_{\text{phones}}} \left(\text{Trg}_j - \text{Out}_j\right)^2. \quad (7)$$

Функционал обучения нейронной сети формулируется на основе функции ошибки так:

$$\sum_{s=1}^{N_{\text{samples}}} E^s = 0,5 \cdot \sum_{s=1}^{N_{\text{samples}}} \sum_{j=1}^{N_{\text{phones}}} \left(\text{Trg}_j - \text{Out}_j\right)^2 \underset{W,B}{\rightarrow} \min, \quad (8)$$

где N_{samples} – количество примеров в обучающем множестве, W – множество весовых коэффициентов, а B – множество коэффициентов смещения всех нейронов сети.

Итак, в процессе обучения необходимо подобрать такие значения весовых коэффициентов и смещений, чтобы минимизировать сумму квадратов отклонений реальных выходов нейронной сети от желаемых ее выходов для всех примеров обучающего множества. Такая задача обучения решается по стратегии обучения с учителем. Одним из наиболее эффективных алгоритмов, решающих эту задачу для многослойной нейронной сети с сигмоидальными функциями активации, есть алгоритм обратного распространения ошибки [16]. В данной статье использован один из улучшенных вариантов этого алгоритма – *Incremental Delta Bar Delta*, позволяющий уменьшить число эпох обучения и с большей эффективностью избегать локальных минимумов функционала обучения [17].

Особенности применения нейросетевого алгоритма распознавания фонем

Распознавание. Процесс распознавания фонем с помощью описанного нейросетевого алгоритма представлен на рис. 4. Прямоугольное окно анализа длиной $T_F = 0,15$ с скользит вдоль речевого сигнала с шагом $T_{\text{shift}} = 0,01$ с. Текущий фрагмент речевого сигнала, вырезаемый этим окном, поступает на вход блока вычисления MFCC-векторов, где подвергается спектральному анализу:

• окно Хэмминга длиной $T_{\text{window}} = 0{,}02$ с, скользящее вдоль текущего речевого фрагмента с шагом $T_{\text{shift}} = 0{,}01$ с, разбивает этот фрагмент на короткие блоки звуковых данных, расположенные последовательно вдоль оси времени;

• для каждого из полученных блоков звуковых данных вычисляется вектор *MFCC*-коэффициентов, или *MFCC*-вектор.

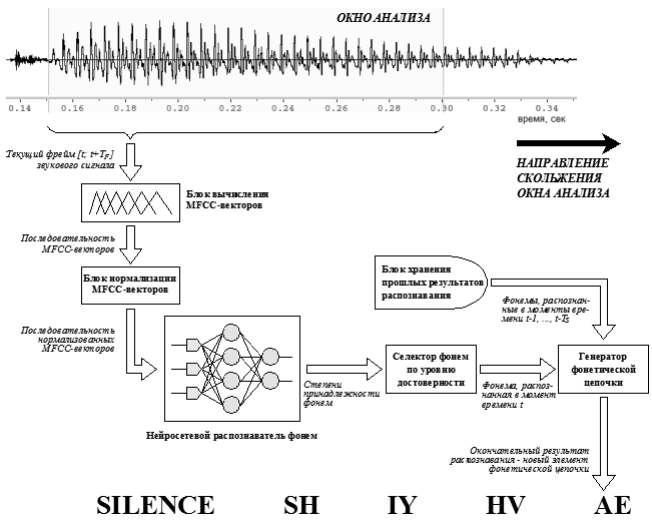

Рис. 4. Структурная схема процесса преобразования звукового сигнала, полученного в результате произнесения английской фразы «*She had your dark suit…*», в фонетическую цепочку с помощью базового варианта нейросетевого алгоритма распознавания фонем

Таким образом, на выходе блока вычисления *MFCC*-векторов получается последовательность *MFCC*-векторов, вычисленных для текущего фрагмента речевого сигнала. Эта последовательность поступает на вход блока нормализации *MFCC*-векторов, в котором каждый из векторов последовательности подвергается нормализации по формуле (1). Соответственно, на выходе блока нормализации *MFCC*-векторов получается последовательность нормализованных *MFCC*-векторов, которая и служит входным сигналом нейросетевого распознавателя фонем.

Нейросетевой распознаватель фонем на основе своего входного сигнала по формуле (6) вычисляет выходной сигнал, который можно трактовать как вектор степеней принадлежности текущего фрагмента речевого сигнала распознаваемым фонемам.

Селектор фонем по уровню достоверности, получая на вход этот вектор в соответствии с формулой (2), определяет текущий результат распознавания j_{res^t} – номер такой фонемы в словаре распознавания, степень принадлежности текущего фрагмента речевого сигнала (t; $t + T_F$) к которой максимальна.

Затем последний блок системы распознавания фонем – формирователь фонетической цепочки – сопоставляет текущий результат распознавания с результатами $j_{\text{res}}^{t-T_{\text{shift}}}, j_{\text{res}}^{t-2\cdot T_{\text{shift}}}, \ldots$ $\ldots, j_{\text{res}}^{t-(N_s-1)\cdot T_{\text{shift}}}$, полученными при анализе предыдущих $N_S - 1$ речевых фрагментов звукового сигнала. Если все эти результаты одинаковы, то генератор фонетической цепочки, выполнив окончательное распознавание фонемы, добавляет в формируемую фонетическую цепочку новый элемент $j_{\text{res}\,t}$-ю фонему из словаря распознавания. В противном случае, когда не все результаты распознавания, полученные на N_S последних шагах анализа, совпадают, решение об окончательном распознавании не принимается, и формируемая фонетическая цепочка остается без изменений.

Описанная процедура сглаживания результатов распознавания, реализуемая генератором фонетической цепочки, нужна для того, чтобы сделать процесс распознавания более устойчивым к случайным ошибкам нейросетевого распознавателя, возможным на переходных участках фонем.

Обучение распознаванию. Процесс преобразования множества звуковых сигналов, их транскрипций и временных меток, задающих границы фонем, в обучающее множество, организуется похожим способом. Прямоугольное окно анализа длиной $T_F = 0{,}15$ с скользит вдоль каждого из обучающих речевых сигналов с шагом $T_{\text{shift}} = 0{,}01$ с. На основе текущего фрагмента речевого сигнала, вырезаемого этим окном, формируется очередной пример (пара *входной сигнал – желаемый выходной сигнал*), добавляемый в обучающее множество. При этом входной сигнал получается в результате мел-частотного кепстрального анализа текущего фрагмента речевого сигнала и последующей нормализации вычисленного мел-частотного кепстрального образа, а желаемый выходной сиг-

нал – в результате анализа фонетической транскрипции и временных меток, задающих границы фонем в речевом сигнале (см. описание входного и желаемого выходного сигналов нейронной сети).

Вариант нейросетевого алгоритма распознавания фонем на основе *bagging*-коллектива нейросетевых распознавателей

Поскольку речь представляет собой нелинейный нестационарный процесс, не удается устойчиво выделить систему речевых признаков, позволяющих проводить абсолютно безошибочную классификацию речевых образов в системе автоматического распознавания. Особенно это проявляется при классификации речевых образов фонем, так как они (в меньшей степени – гласные, в большей – согласные) – весьма нестабильные речевые единицы. При использовании любых систем признаков, даже таких приближенных к механизмам человеческого восприятия, как коэффициенты перцептивного линейного предсказания или применяемые в данной статье мел-частотные кепстральные коэффициенты, признаковое описание одних и тех же фонем может существенно изменяться под влиянием коартикуляции, ускорения темпа речи и других явлений. В данной ситуации повышение достоверности распознавания фонем в устной речи только вследствие увеличения числа признаков и связанного с этим увеличения априорной информации не сможет гарантировать качественное распознавание.

В этих условиях для повышения достоверности распознавания речевых образов логично увеличить количество текущей информации о распознаваемом образе путем объединения отдельных распознавателей в единую систему на принципах коллективного распознавания речи.

Все N распознавателей в такой системе одновременно проводят классификацию поступившего на вход речевого образа. При этом, если разные распознаватели используют один и тот же набор признаков X (например, только мел-частотную кепстрограмму речевого сигнала), как это показано на рис.5,*а*, то объединение их в систему приводит к увеличению количества текущей информации. Если же наборы признаков X_i, $i = 1…N$, индивидуальны для каждого из распознавателей (например, первый распознаватель использует спектрограмму речевого сигнала, второй – мел-частотную кепстрограмму, а третий – набор вейвлет-коэффициентов), как показано на рис. 5,*б*, то в результате объединения таких распознавателей в систему увеличивается количество как текущей, так и априорной информации.

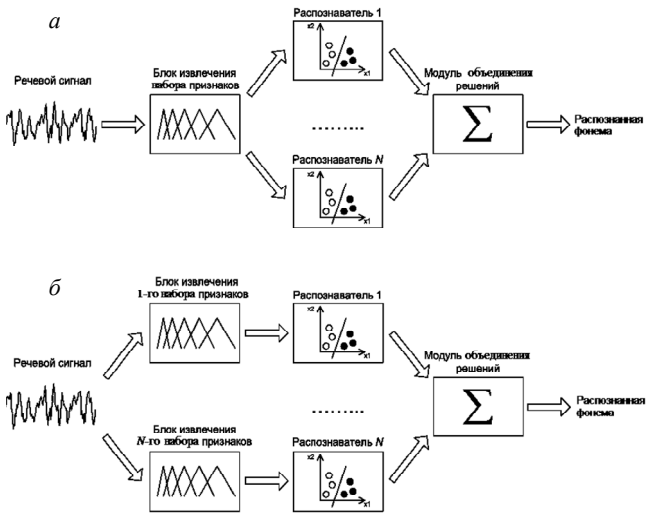

Рис. 5. Объединение отдельных распознавателей в систему: *а* – все распознаватели используют один и тот же набор признаков речевого сигнала; *б* – распознаватели используют разные наборы признаков речевого сигнала

Согласно теории информации [18], в обоих случаях количество информации, перерабатываемой системой при коллективном распознавании, составляет:

$$I(A, X) = H(A) - H(A|X), \qquad (9)$$

где $A = \{A_i\}$, $i = 1…M$ – словарь распознавания (множество распознаваемых слов или фонем); $X=\{X_j\}$, $j = 1…N$ – множество наборов признаков, используемых распознавателями; $H(A)$ – энтропия на входе системы распознавания (исходная энтропия), вычисляемая по формуле

$$H(A) = -\sum_{i=1}^{M} P(A_i) \cdot \log\left(P(A_i)\right); \quad (10)$$

$H(A|X)$ – энтропия на выходе системы распознавания (энтропия решения), которая для коллектива из N распознающих автоматов вычисляется по формуле

$$H(A|X) = H(A|X_1, \ldots, X_N). \qquad (11)$$

Энтропия решения системы коллективного распознавания $H(A|X)$ может уменьшаться при увеличении числа распознавателей в системе, поскольку, согласно [19], условная энтропия с ростом числа фиксируемых условий не возрастает, т.е.

$$H\left(A\middle|X_1,\ldots,X_N\right)\le H\left(A\middle|X_1,\ldots,X_{N-1}\right). \quad (12)$$

При этом строгое равенство наблюдается тогда и только тогда, когда выполняется условие:

$$p\left(A,X_N\mid X_1,\ldots,X_{N-1}\right)= \\ =p\left(A\mid X_1,\ldots,X_{N-1}\right)\times p\left(X_N\mid X_1,\ldots,X_{N-1}\right). \quad (13)$$

Это означает, что при распознавании речи данные, предоставляемые соседними распознавателями коллектива, не дают дополнительной информации к той, которой располагает каждый конкретный распознаватель. Это возможно в двух случаях: когда один из распознавателей всегда принимает безошибочные решения и когда все распознаватели всегда принимают одинаковые решения. Первый случай невозможен на практике, а второй – исключается путем формирования коллектива из различных, а не одинаковых, распознающих автоматов.

Таким образом, возникает задача определения структуры коллектива распознавателей. Существует ряд методов формирования коллектива различных распознающих автоматов [20], среди которых можно выделить три основных:

- *bagging*, или *bootstrap aggregation* – обучение распознавателей на бутстрап-подмножестве базового обучающего множества [21];
- *boosting* – последовательное обучение распознавателей – членов коллектива, при котором каждый следующий распознаватель, включенный в коллектив, обучается так, чтобы компенсировать недостатки всех предыдущих распознавателей [22];
- *mixture of experts* – смесь экспертов, когда в коллектив вводится дополнительный распознаватель, оценивающий компетентность других членов коллектива для каждого входного сигнала и объединяет их индивидуальные решения с учетом вычисленных оценок [23].

Задача распознавания устной речи характеризуется высокой вычислительной сложностью и большими объемами данных для обучения (например, классический речевой корпус для обучения распознаванию английской речи *TIMIT* [24] содержит свыше 500 Мб речевого акустического материала). Для решения такой задачи наиболее целесообразно использование первого подхода – формирования коллектива нейросетевых распознавателей на основе метода *bagging*, так как:

- обучение отдельных нейронных сетей на собственных бутстрап-подмножествах обучающей выборки осуществляется независимо, что позволяет ускорить формирование коллектива путем распараллеливания процессов обучения отдельных нейронных сетей;
- обучающее бутстрап-подмножество может иметь меньший размер, чем базовое обучающее множество, что позволяет ускорить процесс обучения каждой нейронной сети.

Для повышения точности дикторонезависимого распознавания фонем в слитной речи авторами предложен коллективный вариант исходного нейросетевого алгоритма, описанного в начале статьи. В этом варианте решения отдельных членов коллектива объединяются путем равноправного голосования. Формируется коллектив нейронных сетей с помощью метода *bagging*.

Эксперименты

Проведены *две* серии экспериментов по распознаванию фонем устной речи. Целью *первой* серии было сравнение базового и коллективного вариантов нейросетевого алгоритма, а целью *второй* – сравнение классического алгоритма распознавания фонем на основе скрытых Марковских моделей и коллективного варианта нейросетевого алгоритма распознавания. Критерием сравнения была точность распознавания фонем [25].

Материалом для экспериментов послужил классический речевой корпус *TIMIT*, содержащий более пяти часов звукозаписей различных английских фраз, произнесенных 630 дикторами на восьми диалектах американского варианта английского языка. Все звукозаписи имеют временную акустико-фонетическую размет-

ку, выполненную профессиональными фонетистами. Речевой корпус разбит разработчиками на два непересекающихся множества: обучающее и тестовое [24]. Обучение всех алгоритмов распознавания проводилось, соответственно, на обучающем множестве, а оценивание точности распознавания – на тестовом множестве.

Все фонемы, которые встречаются в речевом корпусе *TIMIT*, были сведены к 39 фонетическим классам так, как предложено в [26].

Точность распознавания фонем речевого корпуса TIMIT с помощью нейронных сетей и скрытых Марковских моделей

АЛГОРИТМ РАСПОЗНАВАНИЯ	ТОЧНОСТЬ РАСПОЗНАВАНИЯ, %
Базовый вариант нейросетевого алгоритма	66,80
Коллективный вариант нейросетевого алгоритма из 50 нейросетевых распознавателей	**69,17**
Алгоритм на базе скрытых Марковских моделей	64,21

В ходе первой серии экспериментов одиночный трехслойный персептрон с 230 и 200 нейронами в первом и втором скрытых слоях, обученный на всем обучающем множестве, выполнил распознавание с точностью 66,80 процентов. *Bagging*-коллектив из 50 многослойных персептронов подобной структуры, каждый из которых обучался на *bootstrap*-подмножестве объемом 40 процентов от объема исходного обучающего множества, показал более высокую точность распознавания – 69,17 процента.

В ходе второй серии экспериментов использован распознаватель фонем на базе скрытых Марковских моделей, разработанный в среде HTK [27] с помощью специализированного скрипта [28]. Результаты экспериментов показали, что использование скрытых Марковских моделей позволяет в лучшем случае достигнуть точности распознавания 64,21 процента (лево-правые скрытые Марковские модели для монофонов, 40 гауссовых смесей для моделирования распределения наблюдений).

Заключение. Для решения проблемы диктороннезависимого распознавания фонем в устной речи, возникающей при создании систем автоматического распознавания слов дискретной и слитной речи на основе фонемно-ориентированного метода, авторами предложен алгоритм, основанный на использовании *bagging*-коллектива нейронных сетей типа «многослойный персептрон».

На материале большого речевого корпуса TIMIT экспериментально показано преимущество *bagging*-коллектива нейронных сетей как перед одиночным нейросетевым распознавателем, так и перед распознавателем на базе скрытых Марковских моделей.

Полученные результаты – 69,17 процента правильно распознанных фонем – иллюстрируют конкурентоспособность нейросетевого алгоритма, предложенного авторами, и практическую целесообразность его использования в системах распознавания слитной речи.

1. *Потапова Р.К.* Речевое управление роботом: лингвистика и современные автоматизированные системы. – М.: КомКнига, 2005. – 328 с.
2. *Кодзасов С.В.*, *Кривнова О.Ф.* Общая фонетика. – М.: Рос. гос. гуманит. ун-т, 2001. – 592 с.
3. *Rabiner L.R.* A tutorial on Hidden Markov models and selected application in speech recognition // Proc. of the IEEE. – 1989. – N 77(**2**). – P. 257–286.
4. *Винцюк Т.К.* Анализ, распознавание и интерпретация речевых сигналов. – Киев: Наук. думка, 1987. – 264 с.
5. *Young S.J.* The general use of tying in phoneme-based hmm speech recognisers // Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing (USA), 1. – 1992. – P. 569–572.
6. *Kawai H.*, *Higuchi N.* Recognition of connected digit speech Japanese collected over the telephone networ // Proc. of the 5th Int. Conf. on Spoken Language Processing (Sydney, Australia). – 1998. – P. 341–344.
7. *Ронжин А.Л.*, *Карпов А.А.*, *Ли И.В.* Система автоматического распознавания русской речи *SIRIUS* // Искусственный интеллект. – 2005. – № 3. – С. 590–601.
8. *Пилипенко В.В.* Распознавание ключевых слов в потоке речи при помощи фонетического стенографа // Там же. – 2009. – № 4. – С. 220–224.
9. *Robinson T.*, *Fallside F.* A Recurrent Error Propagation Network Speech Recognition System // Computer Speech & Language. – 1991. – N 5(3). – P. 259–274.
10. *Харламов А.А.*, *Кнеллер Э.Г.* Распознавание ключевых слов в потоке слитной речи на основе нейросетевых технологий // Нейрокомпьютеры: разработка, применение. – 2005. – № 8–9. – С. 88–89.
11. *Галушкин А.И.* Методика решения задач в нейросетевом базисе // Нейронные сети: основы теории. – М.: Горячая линия – Телеком, 2010. – С. 420–466.

12. *Рабинер Л.Р.*, *Шафер Р.В.* Цифровые модели речевых сигналов // Цифровая обработка речевых сигналов. – М.: Радио и связь, 1981. – С. 41–110.
13. *Рабинер Л.Р.*, *Шафер Р.В.* Методы обработки речевых сигналов во временной области // Там же. – С. 110–160.
14. *Аграновский А.В.*, *Леднов Д.А.* Теоретические аспекты алгоритмов обработки и классификации речевых сигналов. – М.: Радио и связь, 2004. – Гл. 1. – 164 с.
15. *Pinkus A*. Approximation theory of the MLP model in neural networks // Acta Numerica. – 199. – N 8. – P. 143–195.
16. *Efficient* BackProp / Y. LeCun, L. Bottou, G. Orr et al. // Neural Networks: Tricks of the trade: Springer Verlag, 1998. – P. 9–50.
17. *Sutton R.S.* Adapting Bias by Gradient Descent: An Incremental Version of Delta-Bar-Delta // Proc. of the Tenth National Conf. on Artificial Intelligence, San Jose, CA, July 1992 / Massachusetts Institute of Technology. – MIT Press, 1992. – P. 171–176.
18. *Барабаш Ю.Л.* Коллективные статистические решения при распознавании. – М.: Радио и связь, 1983. – 224 с.
19. *Файнстейн А*. Основы теории информации. – М.: Изд-во иностр. лит., 1960. – 143 с.
20. *Городецкий В.И.*, *Серебряков С.В*. Методы и алгоритмы коллективного распознавания // Автоматика и телемеханика. – 2008. – № 11. – С. 3–40.
21. *Breiman L.* Bagging Predictors // Machine Learning. – 1996. – 24(**2**). – P. 123–140.
22. *Shrestha D.L.*, *Solomatine D.P*. Experiments with Ada.Boost.RT, an Improved Boosting Scheme for Regression // Neural Computation. – 2006. – N 18(**7**). – P. 1678–1710.
23. *Avnimelech R.*, *Intrator N.* Boosted Mixture of Experts: An Ensemble Learning Scheme // Ibid. – 1999. – N 11(**2**). – P. 483–497.
24. *Zue V.*, *Seneff S.*, *Glass J.* Speech database development at MIT: TIMIT and beyond // Speech Communication. – 1990. – N 9(**4**). – P. 351–356.
25. *Xuedong Huang*, *Alejandro Acero*, *Hsiao-Wuen Hon.* Spoken language processing: a guide to theory, algorithm, and system development. – Prentice-Hall PTR, 2001. – 980 p.
26. *Lee K.-F.*, *Hon H.-W*. Speaker Independent Phone Recognition Using Hidden Markov Models // IEEE Transactions on Acoustics, Speech and Signal Processing. – 1989. – N 37(**11**). – P. 1641–1648.
27. *The HTK* Book (for HTK V. 3.4) / S. Young, G. Evermann, M. Gales et al. // Cambridge University Engineering Department, Cambridge, 2006. – 368 p.
28. *Robinson A.J*. HTK training for TIMIT from Cantab Research, bash shell script, v. 1.3, dawnloaded via – http://www.cantabResearch.com/HTKtimit.html, 2006.

Тел. для справок: +38 0432 59-8413, +38 066 281-9319

●