ОБОЗРЕНИЕ
ПРИКЛАДНОЙ И ПРОМЫШЛЕННОЙ
Том 21 **МАТЕМАТИКИ** **Выпуск 2**
2014

В. П. Ульяненко, **Н. М. Чавчавадзе** (Москва, ООО «Линфо», ТВП). **Редуцирование алфавита при распознавании по матрице рандомизированных оценок фонем.**

Для распознавания речи в [3] был предложен подход, состоящий из следующих этапов: 1) акустический анализ; 2) построение матриц рандомизированных оценок фонем; 3) построение уточненных матриц оценок фонем; 4) восстановление текста.

На первом этапе распознавания непрерывный речевой сигнал представляют в виде последовательности дискретных наблюдений $X_1, X_2, \ldots, X_T$, проводимых в окнах (фреймах) фиксированной длительности в несколько миллисекунд с перекрытиями [1, 2]. Наблюдение X_t $(t = 1, 2, \ldots, T)$ в типичном случае является вектором, координаты которого описывают свойства сигнала, характеризующие соответствие тому или иному элементу фонетического алфавита $O = \{O_1, O_2, \ldots, O_m\}$.

В общем случае одной фонеме может соответствовать некоторое (случайное) число подряд идущих фреймов. Поэтому в ходе этого этапа решается одновременно задача сегментации, т. е. задача первичного определения границ фонем, которая в математическом плане сводится к задаче о разладке. Для этого производят попарное сравнение всех соседних непересекающихся векторов кепстральных коэффициентов [4, 5]. Сходство и различие между классифицируемыми объектами устанавливается в зависимости от выбранной меры близости между ними. Если каждый объект описывается k признаками, то он может быть представлен как точка в k-мерном пространстве. Сходство с другими объектами определяется как соответствующее расстояние. Такой мерой близости может быть, например, расстояние Махаланобиса:

$$\rho_m(X_i, X_j) = (X_i - X_j)^{\mathrm{T}} C^{-1} (X_i - X_j), \tag{1}$$

где: X_i, X_j — координаты i-го и j-го объектов в k-мерном пространстве; C^{-1} — ковариационная матрица генеральной совокупности.

На втором этапе для каждого наблюдения вычисляют правдоподобия того, что это наблюдение соответствует тому или иному элементу алфавита фонем O. Таким образом, по наблюдаемому звуковому сигналу для фонетического алфавита, состоящего из m символов, строится матрица длительностью T фреймов. Для каждого фрейма X_t определяется рандомизированная оценка $\{P(O_t^i),\ i = 1, 2, \ldots, m\}$ истинного значения фонемы во фрейме.

$$\begin{array}{c} O_1^1 \quad O_2^1 \quad \ldots O_t^1 \quad \ldots \quad O_T^1 \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ O_1^i \quad O_2^j \quad \ldots O_t^k \quad \ldots \quad O_T^n \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ O_1^m \quad O_2^m \quad \ldots O_t^m \quad \ldots \quad O_T^m \end{array} \tag{2}$$

где $i, j, k,\ \ n \in 1, 2, \ldots, m,\ \ t \in 1, 2, \ldots T.$

Для полученной матрицы (2) с вероятностями $P(O_t^i)$, в которой одной фонеме может соответствовать некоторое число подряд идущих столбцов, проводят вероятностную оценку границ фонем в исходном сигнале путем попарного сравнения соседних колонок матрицы с помощью статистики Кульбака–Лейблера

$$\rho_{kl}(P(O_t), P(O_{t+1})) = \sum_i P(O_t^i) \ln \frac{P(O_t^i)}{P(O_{t+1}^i)} \tag{3}$$

Далее для каждого определенного временного интервала вычисляют акустическую вероятность фонемы $q(\varphi_t^j)$ и получают решетку фонем с определенными временными интервалами

$$\begin{array}{c} \varphi_1^1,\ \varphi_2^1,\ \ldots,\ \varphi_t^1,\ \ldots,\ \varphi_N^1 \\ \varphi_1^2,\ \varphi_2^2,\ \ldots,\ \varphi_t^2,\ \ldots,\ \varphi_N^2 \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ \varphi_1^m,\ \varphi_2^m,\ \ldots,\ \varphi_t^m,\ \ldots,\ \varphi_N^m \end{array} \tag{4}$$

где m — мощность алфавита фонем, $t \in \{1, 2, \ldots, N\}$, и для каждой фонемы φ_t^j, $t = 1, 2, \ldots, N$, $j = 1, 2, \ldots, m$, вычислена акустическая вероятность q_t^j, т. е. для каждой колонки решетки (4) получают распределение $P_t = \{q_t^j = p(\varphi_t^0 = j), j = 1, 2, \ldots, m\}$.

Экспериментальная апробация такого подхода по определению границ фонем показала, что для русского языка в речевом сигнале образы некоторых фонем фонемного алфавита, состоящего из 58 фонем (ударные и безударные гласные, твердые и мягкие согласные, пауза, шум), близки между собой в смысле расстояния Маланобиса или дивергенции Кульбака–Лейблера. Это приводит к ошибкам статистических критериев, используемых для сегментации оцифрованного речевого сигнала, то есть на этапе выделения фонемомест и построения матрицы (4). В результате матрица (4) содержит вставки и пропуски по сравнению с истинным разбиением речевого сигнала на фонемоместа.

Для повышения надежности сегментации речевого сигнала предлагается объединять некоторые образы фонем в группы по признаку близости их расстояний Махаланобиса внутри групп и высокой различимости между группами. Так, например, оказалось, что полезно объединять в одну группу одноименные гласные (ударные, безударные и во второй степени редукции), одноименные пары твердых или мягких согласных, некоторые пары глухих или звонких согласных и т. д. В результате фонемный алфавит удалось сократить с 58 до 28 фонем.

Одновременно при построении модели фонемного алфавита оказалось полезным создавать динамически по несколько образов одной и той же фонемы в зависимости от диктора, фазирования начала фонемы и т. д.

На третьем этапе с помощью рекурсивных алгоритмов улучшают результаты распознавания, вычислив уточненную матрицу фонем, а на четвертом этапе по полученной решетке фонем с помощью словаря соответствия между транскрибированной и письменной формой слов языка восстанавливают письменный текст.

На третьем и четвертом этапах редуцированный алфавит позволил более чем на порядок понизить сложность вычислений. Кроме того, преимуществом описанного подхода является возможность извлечения информации из исходных данных из существенно меньшего объема речи (ориентировочная оценка — 5 часов речи для языка), что снижает затраты на создание моделей.

СПИСОК ЛИТЕРАТУРЫ

1. *Рабинер Л. Р., Шафер Р. В.* Цифровая обработка речевых сигналов: Пер. с англ./ Под ред. М. В. Назарова и Ю. Н. Прохорова. М.: Радио и связь, 1981.

2. *Young S.*, *Evermann G.*, *Kershaw D.*, *Moore G.*, *Odell J.*, *Ollason D.*, *Valtchev V.*, *Woodland P.* The HTK Book. Cambridge: Cambridge Univ. Engng Dep., 2005.
3. *Чавчавадзе Н. М.*, *Ульяненко В. П.* Способ распознавания речевой звуковой информации. — Обозрение прикл. и промышл. матем., 2012, т. 19, в. 3, с. 473–475.
4. ru.wikipedia.org
5. universal _ ru _ en.academic.ru