УДК 621.391

А.С. Рылов

ДИВЕРГЕНТНЫЙ МЕТОД ОЦЕНКИ ИНФОРМАТИВНОСТИ ПРОСТРАНСТВ ПРИЗНАКОВ ПРИ РАСПОЗНАВАНИИ РЕЧЕВЫХ ОБРАЗОВ

Предлагается метод оценки информативности пространств признаков, позволяющий количественно и качественно оценить состав компонент характеристических векторов, описывающих речевой сигнал на ранней стадии проектирования конкретной системы распознавания речевых образов. На основе предложенного метода получены результаты экспериментальных исследований по оценке информативности различных видов сепстральных параметров для решения задачи текстозависимой верификации личности по речевому сигналу.

Введение

Анализ анатомо-физиологических, психических и социальных факторов, обусловливающих индивидуальные особенности устной речи человека [1–3], выявил факт существования основной и дополнительной информации, заключенной в речевом сигнале. К основной информации относится семантика высказывания (или информация о фонемном составе), к дополнительной – индивидуальность голоса говорящего, а также информация о состоянии человека и его обликовых особенностях. Все это свидетельствует об информационной многозначности речевого сигнала. Выделяя из него различные виды информации, можно решать разнообразные задачи распознавания речевых образов. Распознавание семантической информации относится к распознаванию речи или классификации речевых образов, распознавание же индивидуальности голоса диктора и его личностных особенностей или состояния относится к их идентификации и диагностике соответственно. Таким образом, проблема распознавания речевых образов – это триединство классификационных, идентификационных и диагностических задач. Естественно, что для решения каждой из них пространства признаков речевых сигналов должны качественно отличаться и удовлетворять требованию быть максимально информативными при минимальной размерности.

Так, например, результаты экспериментальных исследований по применению статических и динамических параметров артикуляции при решении различных задач распознавания речевых образов свидетельствуют о том, что параметры, отражающие динамику речевого тракта, более предпочтительны для использования в классификационных задачах, а параметры, отражающие статику, – в идентификационных [4]. Установлено, что надежность распознавания слов от произвольного диктора по Δ-сепстральным параметрам на 5 % выше, чем по сепстральным [5], а надежность верификации диктора по сепстральным параметрам на 5,7 % выше, чем по Δ-сепстральным [6].

Крайне актуальной является задача качественной и количественной оценки размерности характеристических векторов (ХВ) эталонной и тестовой последовательностей. Правильный выбор вида признаков и уменьшение их размерности при сохранении информативности повышает надежность распознавания и быстродействие распознающей системы.

В данной работе предлагается метод оценки информативности признаковых пространств, позволяющий количественно и качественно оценить состав компонент ХВ, выбранных для представления минимальных речевых единиц (МРЕ) на начальной стадии (до разработки классификатора) решения отмеченных выше задач распознавания речевых образов.

1. Теоретическое обоснование метода

Известно, что дивергенция – это оценка степени отличия между двумя классами на основе теории информации [7, 8]. Она обеспечивает возможность ранжировать признаки по спо-

собности разделять классы, т. е. оценивать степень компактности пространств признаков, характеризуя тем самым их информативность.

Пусть два класса образов W_i и W_j имеют нормально распределенные функции плотности вероятности (ФПВ) параметров X_i и X_j соответственно. Тогда дивергенция для них определяется следующим образом:

$$div_{ij} = \frac{1}{2} tr(V_i - V_j)(V_j^{-1} - V_i^{-1}) + \frac{1}{2} tr(V_i^{-1} + V_j^{-1})(M_i - M_j)(M_i - M_j)^T, \quad (1)$$

где $tr|A| = \sum_{i=1}^{p} a_{ij}$ – след матрицы p x p; a_{ij} – диагональные элементы матрицы; V_i и V_j – ковариационные матрицы временных последовательностей параметров X_i и X_j для классов образов W_i и W_j; M_i и M_j – n-мерные векторы-столбцы значений средних компонент n-мерных векторов X_i и X_j. Первое слагаемое правой части (1) без коэффициента ½, т. е.

$$J_{ij} = tr(V_i - V_j)(V_j^{-1} - V_i^{-1}), \quad (2)$$

называют формой дивергенции. Рассмотрим особый случай, когда

$$V_j = V_j = V. \quad (3)$$

Это происходит при совпадении эталонной модели и тестовой реализации. Тогда форма дивергенции (2) становится равной 0, а выражение (1) примет вид

$$\begin{aligned} div_{ij} &= \frac{1}{2} tr\left[V^{-1}(M_i - M_j)(M_i - M_j)^T\right] + \frac{1}{2} tr\left[V^{-1}(M_j - M_i)(M_j - M_i)^T\right] = \\ &= tr\left[V^{-1}(M_i - M_j)(M_i - M_j)^T\right] = \delta^T V^{-1} \delta, \end{aligned} \quad (4)$$

где

$$\delta = M_i - M_j. \quad (5)$$

Таким образом, в случае равенства ковариационных матриц временных последовательностей сепстральных параметров сравниваемых классов образов W_i и W_j, а также при нормальном распределении их ФПВ дивергенция превращается в меру Махаланобиса [9]:

$$d_m^2 = (X - M)^T V^{-1} (X - M). \quad (6)$$

Отличие заключается лишь в том, что в выражении (6) берется разность между текущим значением вектора параметров X и его средним M, а в выражении (4) – между двумя средними (5) сравниваемых классов образов W_i и W_j.

С другой стороны, в случае использования метода сравнения векторно-квантованных (ВК-пространств) признаков дикторов [10] сравниваются центроиды кодовых книг W_i и W_j:

$$cent(C_i) = \frac{1}{\tau_i} \sum_{X \in \tau_i} X(t), \quad (7)$$

где C_i – i-й кластер, состоящий из τ значений векторов параметров X.

Следовательно, центроиды (7) по определению являются средними значениями векторов параметров X. Поэтому дивергенция (4) может быть одновременно мерой близости между двумя ВК-пространствами признаков классов образов и показателем информативности тех или

иных разновидностей ХВ, являющихся центроидами (7). Необходимо отметить, что если параметры X декоррелированы, то дивергенция удовлетворяет принципу аддитивности [7], т. е.

$$div_{ij}(x_1, x_2, ..., x_m) = \sum_{k=1}^{m} div_{ij}(x_k). \quad (8)$$

Таким образом, для оценки информативности ХВ целесообразно применить метод сравнения ВК-пространств признаков при использовании информационной меры близости – дивергенции.

2. Дивергентный критерий информативности ХВ

Для оценки информативности ХВ или компактности пространств признаков распознаваемых образов введен вероятностный критерий, характеризующий вероятность уровня равных ошибок первого и второго рода (EER) и оцениваемый по пересечению кривых распределений вероятностей несовершения ошибок первого и второго рода (кривые 1 и 2 соответственно на рис. 1). Кривая 1 характеризует внутриклассовую вероятность, а кривая 2 – межклассовую. Эти кривые рассчитываются с помощью гистограмм распределения внутриклассовых 1 и межклассовых 2 мер близости (4), показанных на рис. 2.

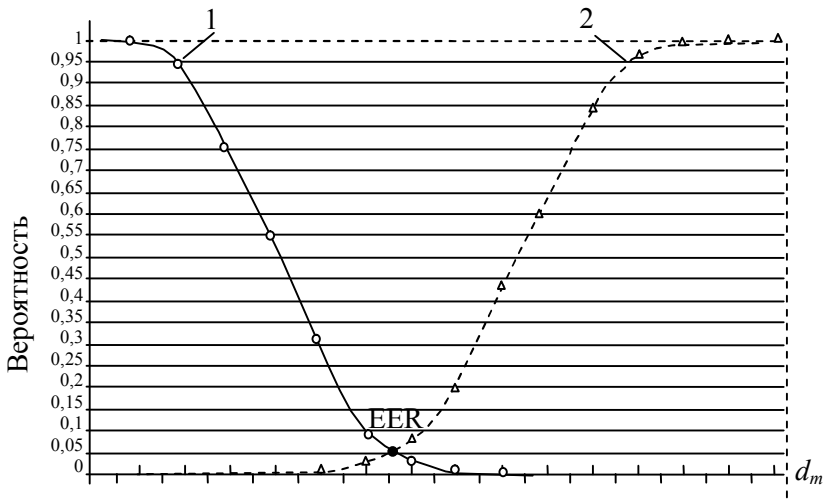

Рис. 1. Распределения вероятностей несовершенных ошибок первого и второго рода

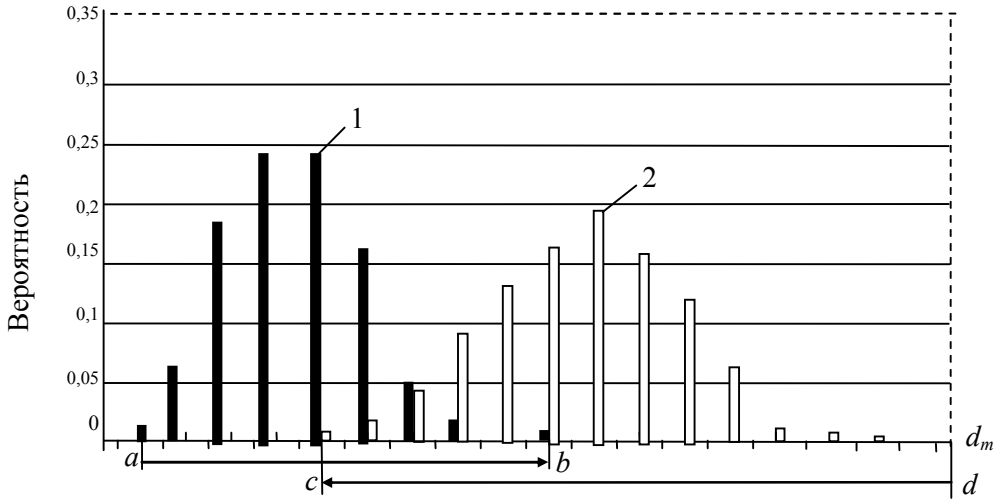

Рис. 2. Гистограммы распределений внутриклассовых и межклассовых мер близости

При этом, как уже отмечалось выше, сравниваются ВК-пространства признаков при использовании соответствующей речевой базы данных. Построение кривых 1 и 2 на рис. 1 осуществляется путем поочередного вычитания текущих значений вероятностей на рис. 2 из предыдущих, начиная с точек a и d соответственно, для которых вероятность берется равной 1. Убывание значений вероятностей на рис. 1 идет в направлениях, указанных стрелками на рис. 2. В табл. 1 приведены результаты расчета гистограмм (первые три строчки) и распределений вероятностей несовершения ошибок первого и второго рода (последние две строчки).

Таблица 1

Результаты расчета гистограмм, EER и распределений вероятностей несовершения ошибок первого и второго рода

d_m	Интервалы	6,404	7,706	9,007	10,31	11,61	12,91	14,21	15,52	16,82	18,12	19,42	20,72	22,02	23,33	24,63	25,93	27,23
Вероятность	внутриклассовая	0,014	0,064	0,186	0,245	0,245	0,164	0,05	0,023	0	0,009	0	0	0	0	0	0	0
	межклассовая	0	0	3E-04	0,002	0,004	0,015	0,043	0,086	0,121	0,163	0,191	0,159	0,113	0,063	0,026	0,01	0,003
	несоверш. ошибок 1-го рода	0	0	0	3E-04	0,002	0,006	0,021	0,064	0,149	0,27	0,434	0,625	0,784	0,897	0,96	0,986	0,996
	несоверш. ошибок 2-го рода	0,995	0,986	0,923	0,736	0,491	0,245	0,082	0,032	0,009	0,009	0	0	0	0	0	0	0

Значение EER на рис. 1 составляет 0,048. Оно характеризует в данном случае информативность Δ-сепстральных (кепстральных) параметров для текстозависимой системы верификации личности по речевому сигналу [11, 12]. Чем меньше значение EER, тем меньше перекрытие между кривыми на рис. 1 и рис. 2 и тем компактнее пространства признаков. Таким образом, критерием компактности является критерий минимума вероятности обобщенной дивергентной ошибки.

3. Расчет гистограмм распределения мер близости в зависимости от разновидностей задач распознавания речевых образов

Изложенный в предыдущем разделе метод позволяет объективно оценивать компактность пространств признаков при решении всех разновидностей задач распознавания речевых образов (классификационных, идентификационных и диагностических). Отличие в оценивании будет заключаться лишь в том, что из себя будут представлять классы образов W_i и W_j при сравнении их ВК-пространств признаков с помощью мер (4), по значениям которых будут строиться гистограммы распределений на рис. 2.

Так, например, при решении задачи распознавания речи с существенными ограничениями на канал и число пользователей [3] в качестве W_i и W_j для построения гистограммы распределения внутриклассовых мер близости (см. гистограмму 1 на рис. 2) следует брать одинаковые речевые единицы, произносимые одним и тем же диктором. При расчете гистограммы распределения межклассовых мер близости (см. гистограмму 2 на рис. 2) в качестве W_i и W_j надо брать отличающиеся речевые единицы, произносимые одним и тем же диктором.

При распознавании речи без существенных ограничений на канал и число пользователей для расчета кривой 1 на рис. 2 в качестве W_i и W_j необходимо брать одинаковые МРЕ из заданного словаря, произнесенные множеством дикторов, а для расчета кривой 2 на рис. 2 в качестве W_i и W_j должны использоваться отличающиеся МРЕ из словаря, произнесенные этим же множеством дикторов.

Для идентификационных задач распознавания речевых образов при расчете кривой 1 на рис. 2 в качестве W_i и W_j могут использоваться две одинаковые фразы, произносимые одним диктором, а для расчета кривой 2 на рис. 2 – эти же фразы, произносимые разными дикторами, и т. п. Во всех случаях должна использоваться представительная база данных, состоящая из соответствующих единиц речи, произнесенных множеством дикторов.

При этом следует соблюдать два требования. Во-первых, МРЕ в базе должны подбираться так, чтобы каждая из них состояла из одинакового количества отличающихся стационарных участков фонем T_s и переходов между ними T_m. Таким образом создаются условия для выбора инвариантной размерности кодовых книг в процедуре векторного квантования для всех речевых единиц в базе. Во-вторых, размерность I кодовых книг должна выбираться в соответствии с равенством

$$I = T_s + T_m . \tag{9}$$

Тогда каждый центроид в кодовой книге будет представлять собой либо стационарный, либо переходной участок между фонемами в МРЕ.

4. Результаты экспериментальных исследований по оценке информативности пространств признаков, выполненных с помощью дивергентного метода

С помощью изложенного выше метода были проведены исследования [8] по оценке информативности пяти разновидностей сепстральных параметров, используемых при решении идентификационных задач распознавания речевых образов (табл. 2).

Таблица 2

Разновидности сепстральных параметров

№ п/п	Формульная запись	Название	Сокращение
1	$c_{FL}(n) = \frac{2}{N}\sum_{k=1}^{N} \log\left\|Y_k\right\|^2 \cos\frac{k\pi n}{N}$, где $n = 1, 2, \ldots, N$, $k = 1, 2, \ldots, N$, Y_k – значения энергий в полосах гребенки перекрывающихся полосных фильтров с центральными частотами k	Сепстральные параметры, рассчитываемые с помощью обратного преобразования Фурье на линейной частотной шкале в герцах	ФЛСК
2	$c_{LP}(n) = \sum_{k=1}^{n-1}\left(1-\frac{k}{n}\right)a_k \cdot C(n-k) + a_n$, где a – параметры предсказания; p – порядок модели ЛП, $1 < n \le p$	Сепстральные параметры, рассчитываемые из коэффициентов линейного предсказания	ЛПСК
3	$c_{DKL}(n) = \frac{2}{N}\sum_{k=1}^{N} \log\left\|Y\right\|^2 \cos\frac{(2k+1)n\pi}{2N}$, где $k = 1, 2, \ldots, M$, $n = 1, 2, \ldots$	Сепстральные параметры, рассчитываемые с помощью дискретно-косинусного преобразования на линейной частотной шкале в герцах	ДКЛСК
4	$c_{DKM}(n) = \sum_{k=1}^{N} \log\left\|Y\left(i \equiv f_{гц}(\theta)\right)\right\| \cos\frac{(2k-1)n\pi}{2N}$, где $n = 1, 2, ..., k = 1, 2, ..., N$, $f_{гц}(\theta)$ – зависимость, связывающая частоту в герцах и частоту в мелах	Сепстральные параметры, рассчитываемые с помощью дискретно-косинусного преобразования на частотной шкале в мелах	ДКМСК
5	$c_{FM}(n) = \frac{2}{N}\sum_{k=1}^{N} \log\left\|Y\left(i \equiv f_{гц}(\theta)\right)\right\| \cos\frac{k\pi n}{N}$, где $n = 1, 2, \ldots, k = 1, 2, \ldots, N$	Сепстральные параметры, рассчитываемые с помощью обратного преобразования Фурье на частотной шкале в	ФМСК

		мелах	

Более детальное описание этих разновидностей сепстральных параметров изложено в работе [13], в которой была использована база данных, состоящая из произношений одной и той же фразы двадцатью дикторами, причем каждый из них произносил ее пять раз. В табл. 3 приведены результаты этих исследований.

Внутри каждой из пяти разновидностей сепстральных параметров исследования проводились для трех типов ХВ, состоящих только из c- или Δc -параметров, а также из их совместного использования ($c + \Delta c$). Здесь c – это сепстральные параметры статики, а Δc – параметры динамики речевого тракта [4]. Кроме того, все перечисленные эксперименты выполнялись для разного количества сепстральных параметров (от 19 до 21).

Таблица 3

Результаты тестирования по оценке информативности сепстральных параметров для решения идентификационных задач

Номер эксп.	Тип сепстрального коэффициента	Количество сепстральных коэффициентов	EER		
			c	$c + \Delta c$	Δc
1	ДКЛСК	19	0,058052	**0,048268**	0,089740
2		20	0,067284	0,052491	0,106129
3		21	0,072972	0,052315	0,111927
4	ФЛСК	19	0,054610	0,058766	0,105455
5		20	0,062787	0,065498	0,101323
6		21	0,074493	0,067681	0,104762
7	ЛПСК	14	0,066883	**0,048030**	0,133268
8	ДКМСК	19	0,086277	0,090062	0,130065
9		20	0,071991	0,078225	0,121580
10		21	0,078333	0,064675	0,095455
11	ФМСК	19	0,069156	0,069242	0,117273
12		20	0,059524	0,061732	0,107879
13		21	0,059697	0,056126	0,090346

Таким образом, анализ результатов исследований, представленных в табл. 3, дает основание утверждать следующее:

1. Самые минимальные значения (EER = 0,048030 и EER = 0,048268) соответствуют ЛПСК и ДКЛСК (эксперименты № 7 и № 1). При этом ХВ должны включать параметры статики и динамики ($c + \Delta c$). Следует также отметить, что c -параметров для ЛПСК было всего 14 вместо 19 в ДКЛСК.

2. Увеличение количества сепстральных коэффициентов от 19 до 21 для разновидностей, рассчитываемых на линейной частотной шкале в герцах, ведет к увеличению EER для всех типов ХВ (c , Δc , $c + \Delta c$), а для разновидностей, рассчитываемых на линейной частотной шкале в мелах, – к уменьшению EER. Однако это уменьшение не является столь существенным, чтобы отдать предпочтение использованию ДКМСК или ФМСК.

3. В отличие от классификационных задач, для решения идентификационных более эффективными являются ДКЛСК, чем ДКМСК, причем это справедливо для всех трех типов ХВ. Таким образом, выбор типа частотной шкалы для расчета сепстральных коэффициентов является принципиальной отличительной особенностью при формировании ХВ для решения классификационных и идентификационных задач распознавания речевых образов. Это утверждение основано на том, что установлена [13] устойчивая обратная закономерность – преимущество ДКМСК и ФМСК по сравнению с ДКЛСК и ФЛСК при решении классификационных задач распознавания речевых образов.

Заключение

Как правило, при оценке информативности тех или иных видов параметров для решения определенных задач распознавания речевых образов используют конкретные системы. В каждой из них применен свой метод распознавания. Поэтому говорить об объективности этих оценок не всегда представляется возможным. Изложенный в данной работе метод дивергентной оценки информативности пространств признаков является инструментарием для получения объективных оценок, так как они выполняются до стадии разработки самого классификатора и поэтому не зависят от качества самих алгоритмов распознавания. При этом метод является универсальным, он позволяет оценивать компактность пространств признаков при решении всех разновидностей задач (классификационных, идентификационных и диагностических) по критерию минимума вероятности обобщенной дивергентной ошибки.

С помощью предложенного дивергентного метода оценки информативности пространств признаков получены следующие основные результаты:

1. Установлено, что в отличие от классификационных задач распознавания речевых образов, для решения идентификационных задач сепстральные коэффициенты должны рассчитываться на линейной частотной шкале в герцах.

2. Наиболее эффективными видами сепстральных коэффициентов для решения идентификационных задач являются коэффициенты, рассчитываемые по параметрам линейного предсказания и с помощью дискретно-косинусного преобразования на линейной частотной шкале в герцах.

3. Использование на практике установленных выше закономерностей позволит повысить надежность систем распознавания личности по речевому сигналу.

Список литературы

1. Галунов, В.И. Речь, эмоции и личность: проблемы и перспективы / В.И. Галунов // Речь, эмоции и личность: материалы и сообщ. Всесоюзн. симп. – Л., 1978. – С. 3–12.

2. Рылов, А.С. Анализ речи в распознающих системах / А.С. Рылов. – Минск: Бестпринт, 2003. – 263 с.

3. Рылов, А.С. Три основные задачи проблемы распознавания речевых образов / А.С. Рылов // Весці Нац. акад. навук Беларусі. Сер. фіз.-тэхн. навук. – 2000. – № 2. – С. 100–114.

4. Рылов А.С. Теоретические основы формирования компактных пространств признаков для решения задач автоматического распознавания речевых образов / А.С. Рылов // Весці Нац. акад. навук Беларусі. Сер. фіз.-тэхн. навук. – 2004. – № 4. – С. 95–105.

5. Nouza, J. On the speech feature selection problem: are dynamic features more important then the static ones? / J. Nouza // Speech communication and technology (Eurospeech-95): Proc. of European conf. / ESCA. – Madrid, 1995. – Vol. 2. – P. 919–922.

6. Soong, F.K. On the use instantaneous and transititional spectral information in speaker recognition / F.K. Soong, A.E Rosenberg // IEEE Trans. on ASSP. – 1988. – Vol. 36, № 6. – P. 71–879.

7. Campbell, J.P. Speaker recognition: tutorial / J.P. Campbell // Proc. of IEEE. – 1997. – Vol. 85, № 9. – P. 1437–1462.

8. Рылов, А.С. Исследование сепстральных параметров для решения идентификационных задач распознавания речевых образов / А.С. Рылов, В.А. Чижденко, Т.В. Левковская // Доклады БГУИР. – 2004. – № 6. – С. 39–45.

9. Распознавание слуховых образов / Н.Г. Загоруйко [и др.]. – Новосибирск: Наука, 1970. – 383 с.

10. Рылов, А.С. О распознавании личности по ВК-пространству речевых признаков / А.С. Рылов // Доклады НАН Беларуси. – 2004. – Т. 48, № 4. – С. 38–41.

11. Рылов, А.С. Способ распознавания речевых образов преимущественно для текстозависимой верификации диктора по речевому сигналу и устройство для его осуществления: заявка a20030890 на патент с приоритетом от 23.09.2003. / А.С. Рылов, В.А. Чижденко // Опубл. БИ. – № 3. – 2004.

12. Рылов, А.С. Речевая текстозависимая система верификации личности на основе сравнения векторно-квантованных пространств признаков / А.С. Рылов, В.К. Конопелько, В.А. Чижденко // Доклады БГУИР. – 2005. – № 6. – С. 88–96.

13. Рылов, А.С. Исследование сепстральных методов формирования характеристических векторов для систем распознавания речи / А.С. Рылов, Т.В. Левковская, В.А. Чижденко // Доклады БГУИР. – 2004. – № 6. – С. 32–38.

Поступила 07.03.06

Белорусский государственный университет
информатики и радиоэлектроники,
Минск, П. Бровки, 6
e-mail: kafsiut@bsuir.unibel.by

A.S. Rylov

DIVERGENT ESTIMATION METHOD OF FEATURE SPACE INFORMATIVENESS FOR IMAGE SPEECH RECOGNITION

An estimation method of feature space informativeness, allowing quantitative and qualitative evaluations of characteristic vector components, which describe the speech signal at the early speech recognition stage is suggested. The experimental results are obtained on the basis of the proposed method. They are applied to informativeness estimation of different cepstral features used for text dependent speaker verification by a speech signal.