УДК 004.934.2

В.Н. Сорокин, А.А. Тананыкин

Распознавание пола диктора с помощью метода Парзена

Распознавание пола диктора выполнялось в пространстве параметров модели голосового источника, найденных путем решения обратной задачи. Исследования проводились на базе данных, состоящей из русскоязычных числительных, произнесённых в обычных условиях. Функция плотности вероятности для каждого пола оценивалась методом Парзена с гауссовым ядром. Распознавание пола производилось по максимуму правдоподобия Байеса. Ошибка распознавания пола на сегментах ударных гласных не превышает 2%.

Ключевые слова: распознавание пола диктора, метод Парзена, ядерная оценка плотности вероятности, голосовой источник.

Определение пола диктора играет важную роль при распознавании речи, подтверждении личности диктора по его голосу или идентификации диктора, поскольку распознавание пола диктора позволяет существенно сузить область принимаемых признаками значений.

В [1] сообщается о безошибочном распознавании пола диктора, однако для этого требовалось несколько секунд речевого сигнала. При сокращении интервала анализа до 500 мс число ошибок возрастает, например, по данным [8], до 7%, а при интервале в 20 мс – до 10%.

Распознавание пола по долговременному спектру вносит задержку в принятие решения о поле диктора, и эта задержка далеко не всегда приемлема в технических приложениях. Цель данной работы состоит в определении возможности распознавания пола диктора при помощи метода Парзена на сравнительно коротком сегменте речи.

Для определения параметров голосового источника может быть достаточно одного периода основного тона, длина которого не превышает 20 мс. На основе полученных параметров формируется пространство признаков, в котором при помощи метода Парзена оцениваются функции плотности вероятности для мужского и женского пола. Вероятность ошибочного распознавания пола определяется путём интегрирования под функцией минимума этих двух функций.

Описание параметров распознавания. Технология оценки параметров модели голосового источника, принятая в данной работе, была описана в [2]. В этой работе для получения устойчивого решения используется модель (1) формы импульса площади голосовой щели, детально описанная в [4]. В модели (1) символами t обозначены относительные значения времен, а символами T – абсолютные.

$$S(t)=\begin{cases} S_{\max}\left[\sin\left(\dfrac{\pi t}{2t_1T_0}\right)\right]^p, & 0\le t\le t_1T_0, \\ S_{\max}\left[\cos\left(\dfrac{\pi(t-t_1T_0)}{2(t_2-t_1)T_0}\right)\right]^q, & t_1T_0<t\le t_2T_0, \\ 0, & t_2T<t\le T_0, \end{cases} \tag{1}$$

где $S_{\max}$ – максимальная площадь открытия голосовой щели ($S_{\max}$ = 0,2 см^2); t_1T_0 – момент максимального открытия голосовой щели ($T_{S\max}$); t_2T_0 – момент закрытия голосовой щели (T_{clos}), а p и q – коэффициенты, определяющие скорость раскрытия и закрытия голосовой щели; T_0 – период основного тона (мс). Импульс, описываемый этой моделью, представлен на рис. 1.

В проводимых исследованиях использовались параметры этой модели, а также некоторые временные параметры производной по времени от функции изменения площади голосовой щели. Исследуемые параметры представлены в табл. 1. В экспериментах по распознаванию пола диктора коэффициенты p и q оказались мало информативными.

Оценка плотности вероятности. В настоящей работе распознавание пола диктора выполнялось с использованием метода Парзена для аппроксимации функции многомерной плотности вероятности в пространстве параметров голосового источника. Метод Парзена относится к классу ядерных непараметрических оценок плотности распределения вероятности [7].

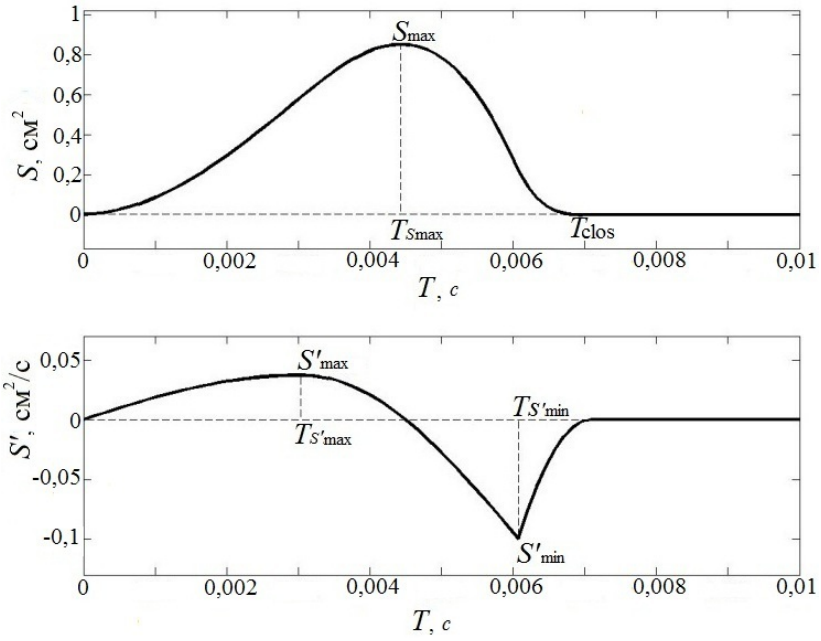

Рис. 1. Площадь голосовой щели (вверху) и её производная

Таблица 1

Исследуемые параметры

№	Параметр	Обозначение
1	Относительный момент времени для максимального значения площади	$t_1 = T_{S\max}/T_0$
2	Относительный момент времени закрытия голосовой щели	$t_2 = T_{clos}/T_0$
3	Период основного тона	T_0
4	Относительный интервал между моментом максимума площади голосовой щели $T_{S\max}/T_0$ и моментом максимума первой производной от площади голосовой щели $T_{S'\max}/T_0$	$\Delta t_1 = \frac{T_{S\max} - T_{S'\max}}{T_0}$
5	Относительный интервал между моментом минимума производной от площади $T_{S'\min}/T_0$ и моментом максимума площади голосовой щели $T_{S\max}/T_0$	$\Delta t_2 = \frac{T_{S'\min} - T_{S\max}}{T_0}$
6	Значение минимума первой производной от площади голосовой щели	$S'_{\min}(t)$

В этот метод можно вложить физический смысл, который заключается в том, что в окрестности каждого вектора, принадлежащего некоторому классу объектов, находятся векторы, также принадлежащие этому классу, причем вероятность появления этих векторов убывает по мере удаления от исходного вектора. Естественно принять закон убывания в виде нормального распределения. Оценка функции плотности вероятности $P(X)$ в многомерном случае по методу Парзена с гауссовым ядром [7] представляется как

$$P(X) = \frac{1}{N}\sum_{i=1}^{N} \frac{1}{(2\pi)^{d/2} h^d |\Sigma|^{1/2}} \exp\left[-\frac{(X - X_i)\Sigma(X - X_i)}{2h^2}\right], \quad (2)$$

где X_i – независимые и одинаково распределённые наблюдения некоторой случайной величины; N – размер выборки; h – некоторый положительный параметр, являющийся функцией от числа наблюдений N; $\sum$ – ковариационная матрица. Поскольку плотность вероятности восстанавливается в малой окрестности каждого вектора из обучающей выборки, то можно принять, что ковариационная матрица $\sum$ является диагональной.

Для гауссова ядра в этом методе единственным неизвестным параметром для каждого распределения является его дисперсия. Эта дисперсия находилась следующим образом. Для каждой пары векторов в обучающей выборке вычислялось расстояние в евклидовой метрике и затем находилось среднее минимальное расстояние $\bar{r}$ для всей выборки. Среднеквадратичное отклонение σ в каждом локальном нормальном распределении назначалось как $\sigma = 1{,}2 \cdot \bar{r}$.

Существенным недостатком оценки по методу Парзена является быстрое возрастание количества вычислений с увеличением числа узлов сетки, по которой проводится интегрирование, что может вынудить ограничивать точность вычислений. Здесь приходится работать в маломерном пространстве признаков.

База речевых данных. Эксперименты проводились на базе речевых данных, описанной в [2]. Эта база состоит из числительных русского языка от 0 до 9 для 49 мужчин и 37 женщин, определенных путем кластеризации в пространстве формантных частот как характерные представители из множества голосов для 243 мужчин и 186 женщин. Каждый диктор произнес от 400 до 800 слов через 4 типа микрофонов, расположенных на разном расстоянии от диктора. В распознавании пола участвовало 26404 произнесения слов мужчинами и 15109 произнесений слов женщинами. К анализу допускались только слова с отношением сигнал/шум не ниже +10 дБ.

Распознавание пола диктора выполнялось на сегментах ударных гласных длительностью от 70 до 200 мс. Эти сегменты были найдены специальным алгоритмом, описанным в [3].

Обсуждение. Использование относительных значений временных параметров позволяет в значительной степени ликвидировать их зависимость от периода основного тона. Ошибка определения пола для каждого отдельно взятого параметра представлена в табл. 2.

Таблица 2

Вероятность ошибочного распознавания пола по параметрам голосового источника

Параметр	t_1	t_2	T_0	Δt_1	Δt_2	$S'_{\min}(t)$
Вероятность ошибки, %	20,33	32,31	19,11	18,01	60,58	19,53

В исследованиях по непосредственному измерению воздушного потока через голосовую щель [5, 6] было найдено, что отношение длительности интервала открытой голосовой щели к периоду основного тона у женщин больше, чем у мужчин. Это вполне согласуется с полученными результатами, поскольку, из табл. 2 видно, что параметр t_2 обладает неплохими разделительными свойствами. Однако ошибка распознавания пола по параметру t_2 в полтора раза выше, чем по параметру t_1. Иначе говоря, момент достижения функцией площади голосовой щели максимума более информативен для различения пола, чем момент закрытия голосовой щели. При рассмотрении двумерных подпространств признаков можно заметить схожую ситуацию. Ошибка распознавания пола в пространстве (t_2, T_0) оказалась близкой к 8,4%, тогда как ошибка в пространстве (t_1, T_0) была заметно ниже – около 5,7%.

В трехмерных подпространствах ошибки распознавания пола значительно ниже, чем в одномерных или двумерных подпространствах, причем в четырех подпространствах ошибка ниже 3%, а в двух – ниже 2% (табл. 3).

Таблица 3

Вероятность ошибочного распознавания пола в трехмерных подпространствах

Параметр	(t_1, t_2, T_0)	$(t_1, t_2, \Delta t_1)$	$(t_1, t_2, S'_{\min}(t))$	$(t_1, T_0, \Delta t_1)$	$(t_1, T_0, S'_{\min}(t))$	$(t_2, T_0, \Delta t_1)$
Вероятность ошибки, %	2,71	2,76	2,97	1,58	1,80	2,37

По сравнению с работой [2], в которой была получена суммарная ошибка около 10% на той же базе данных, в настоящей работе ошибки снижены почти в 6 раз. Такое снижение ошибок идентификации было достигнуто как за счет предварительной сортировки данных, выпадающих из распределения вероятностей, так и вследствие более точного восстановления плотности вероятностей методом Парзена. Это весьма хорошая оценка, на основании которой можно надеяться на идентификацию пола диктора на относительно коротком сегменте речевого сигнала с голосовым возбуждением.

В силу нестабильности оценки параметров голосового источника с использованием метода обратной фильтрации на каком-то речевом сегменте может произойти отказ от анализа параметров голосового источника. Это приведет к задержке решения о поле диктора до тех пор, пока не появится сегмент с хорошим качеством анализа. Поэтому для идентификации пола целесообразно привлечь дополнительную информацию, например формантные частоты.

Если на некотором речевом сегменте получены решения и о параметрах голосового источника, и о формантных частотах, то достигается минимальная вероятность ошибки распознавания с достаточно хорошим запасом надежности. Однако такое решение может быть получено далеко не для

всех речевых сегментов. В других случаях может присутствовать информация только о голосовом источнике или только о формантных частотах.

Заключение. Применение метода Парзена в задаче распознавания пола диктора позволило получить вероятность ошибки распознавания до 2%, что в несколько раз ниже результата, полученного на этой же базе данных в [2]. Несмотря на то, что уже известен факт безошибочного распознавания пола [1] диктора, предложенный подход позволяет определить пол по значительно меньшей длительности сигнала.

Литература

1. Ромашкин Ю.Н. Распознавание пола диктора на основе GMM-модели голоса / Ю.Н. Ромашкин, Ю.О. Петров // Речевые технологии. – 2009. – №. 2. – С. 31–38.

2. Сорокин В.Н. Распознавание пола диктора по голосу / В.Н. Сорокин, И.С. Макаров // Акустический журнал. – 2008. – Т. 54, № 4. – С. 1–9.

3. Сорокин В.Н. Сегментация и распознавание гласных / В.Н. Сорокин, А.И. Цыплихин // Информационные процессы. – 2004. – Т. 4, № 2. – С. 202–220.

4. Doddington G.R., A computer method of speaker verification: Ph.D. thesis / Department of Electrical Engineering. – USA, Madison: University of Wisconsin, 1970.

5. Glottal airflow and transglottal air measurements for male and female speakers in low, normal, and high pitch / E.B. Holmberg, R.E. Hillman, J.S. Perkell // J. Voice. – 1989. – Vol. 4. – С. 511–529.

6. Comparison among aerodynamic, electroglottographic, and acoustic spectral measures of female voice / E.B. Holmberg, R.E. Hillman, J.S. Perkell et al. // J. Speech Hear. Res. – 1995. – Vol. 38. – P. 511–529.

7. Parzen E. An estimation of a probability density function and mode // Ann.Math.Stat. – 1962. – Т. 33. – P. 1065–1076.

8. Sigmund M. Gender distinction using short segments of speech signal // Int. J. of Computer Science and Network Security. –2008. – Vol. 8, №. 4. – P. 159–163.

Сорокин Виктор Николаевич
Д-р физ.-мат. наук, вед. науч. сотрудник Институт проблем передачи информации (ИППИ) им. А.А. Харкевича (РАН), Москва
Тел.: (495) 699-50-96
Эл. почта: vns@iitp.ru

Тананыкин Александр Александрович
Научный сотрудник ИППИ
Тел.: 8-916-394-88-64
Эл. почта: tananykin@mail.ru

Sorokin V.N., Tananykin A.A.
Speaker gender recognition by Parzen method

This paper presents a gender recognition study, which is based on vocal source area parameters and Parzen method. The vocal slit area parameters are derived by inverse filtering method. The probability – density function for each gender was estimated by Parzen method with common Gaussian kernel. In experiments we used a database of Russian digits recorded in comfortable conditions with several types of microphones. Clustering from the original database method selected 86 speakers, including 49 men and 37 women representing the diversity of voices. This database was used to estimate the dynamics parameters of the glottis. Obtained total recognition error rate for short vowels was about 2%.

Keywords: gender recognition, Parzen method, kernel density estimation.