Распознавание пола по параметрам голосового источника

Сорокин В.Н., *доктор физико-математических наук, с.н.с.,*

Ромашкин Ю.Н., *кандидат технических наук,*

Тананыкин А.А., *аспирант*

Распознавание пола диктора выполняется в пространстве параметров голосового источника, найденных путём решения обратной задачи, и в пространстве первых трёх формантных частот на ударных гласных. Эти пространства практически независимы. Вероятность ошибки распознавания пола в пространстве параметров голосового источника ниже 2%, а в пространстве формантных частот близка к нулю. Установлено, что длительность речевого сигнала, необходимого для принятия решения о поле диктора, может варьироваться от нескольких десятков миллисекунд до нескольких секунд.

• распознавание пола диктора • параметры голосового источника • формантные частоты • вероятность ошибки

Recognition of the speaker gender is performed in the space of parameters of the voice source, found by solving the inverse task, and in the space of the first three formant frequencies of the stressed vowels. It is shown that these spaces are independent. The probability of recognition error in voice source space is less than 2% and in the formant space is close to zero. It is established that duration of the speech signal, necessary for speaker gender decision-making, may vary from a few tens of milliseconds to a few seconds.

• recognition of the speaker gender • parameters of the voice source • formant frequencies • probability of error rate

Введение

Распознавание пола диктора по голосу играет существенную роль в задачах автоматического распознавании его речи и идентификации. Область существования информативных признаков речи для каждого конкретного диктора существенно уже, чем диапазон значений этих признаков для всего множества дикторов. Поэтому для снижения ошибки автоматического распознавания речи желательно предварительно установить тип диктора или, по крайней мере, его пол. При автоматической идентификации диктора выполняется последовательное сравнение параметров произнесённого звукосочетания с аналогичными параметрами всех дикторов, имеющимися в базе данных. Диктор идентифицируется по критерию максимального правдоподобия принадлежности этих параметров к одному из множеств в базе данных. Предварительное определение пола диктора не только позволяет сократить объём перебора данных, но и указывает на методы анализа, специфические для мужчин и женщин.

Физической основой для различения мужских и женских голосов служат различия в анатомических размерах и плотности и упругости тканей речевого тракта. Различие в анатоми-

ческих размерах тракта непосредственно сказывается на распределении резонансных частот для одного и того же звука, произнесённого мужчиной или женщиной. Косвенно это различие проявляется в форме спектра соответствующих звуков, на которую влияют и параметры импульса голосового возбуждения. Параметры тканей сказываются на ширине резонансов тракта и также влияют на форму спектра. Существенная информация содержится в скорости речи и длительности речевых сегментов.

Интуитивно очевидным признаком, отделяющим мужские голоса от женских, является частота основного тона F_0. Этот признак используется во многих системах распознавания пола. Однако, как было показано в [1] на базе голосов более чем 400 дикторов, вероятностные распределения F_0 мужских и женских голосов заметно пересекаются даже в сравнительно идеализированных условиях произнесения отдельных слов. Суммарная ошибка распознавания пола в этих экспериментах составила около 18–21%.

В произвольной слитной речи на частоту основного тона влияет фразовая интонация, сдвигающая преимущественно F_0 мужских голосов вверх в область частот женских голосов. При этом различительная способность частоты основного тона падает, в основном, для мужских голосов.

Непосредственные измерения импульсов голосового источника, выполненные в [2–4], установили, что женские голоса, в отличие от мужских, часто не имеют интервала закрытой голосовой щели. Это дало основание для применения анализа источника голосового возбуждения для распознавания пола [5–7]. Использование параметров модели голосового источника в [1] позволила снизить суммарную ошибку автоматического распознавания пола до 10%.

Совместное влияние голосового источника и анатомии речевого тракта проявляется в некоторых параметрах спектра. В [8–9] было найдено, что соотношение между амплитудами первой и второй гармоник основного тона в спектре и амплитудами первой и третьей форманты различно для мужских и женских голосов. Помимо частоты основного тона, амплитуды первой и второй гармоник спектра, частота третьей форманты также содержит информацию о поле диктора [10].

Наиболее технологичный подход к использованию спектральных характеристик состоит в переходе к кепстральному представлению в шкале мел-частот. При вычислении кепстра обычно используется долговременный спектр на интервале времени до нескольких секунд. Увеличение числа кепстральных коэффициентов мало сказывается на эффективности дискриминации пола диктора в благоприятных акустических условиях, но существенно влияет на устойчивость к шумам. Кепстральный анализ достаточно устойчив и к влиянию сжатия речи. В [11] сообщается о безошибочном распознавании пола диктора на базе данных 90 мужчин и 30 женщин, речь которых подвергалась сжатию с помощью стандартного GSM-кодека. При этом интервал анализа составлял 5–10 секунд. При сокращении интервала анализа до 500 мс число ошибок возрастает, например, по данным [12], до 7%, а при интервале в 20 мс — до 10%.

На форму спектра, в частности на его наклон, влияет эффект Ломбарда, возникающий при разговоре в условиях повышенного акустического шума. При этом не только увеличивается уровень громкости речи, возрастает частота основного тона, но и меняются характеристики импульсов голосового возбуждения с повышением уровня высокочастотных компонент. Амплитудно-частотные характеристики акустического и телефонного канала

связи также искажают спектр речи. Особенно заметно влияние реверберации помещений [1].

Формантные частоты гласных звуков определяются как анатомическими размерами тракта, так и артикуляцией конкретных звуков. Эти частоты содержат информацию о поле диктора [13]. Субъективная вероятность распознавания пола диктора по фонемам составляет более 98% [14]. В [7] сообщается, что различительная способность частоты второй форманты даже выше, чем частоты основного тона. Частоты третьей и четвёртой формант несут больше информации о поле диктора, чем частоты первой и второй форманты [15]. В условиях, когда принудительно фиксируется частота основного тона и длина речевого тракта, восприятие пола диктора существенно затрудняется [16].

Установлено, что наилучшие результаты получаются при анализе сегментов речи, содержащих гласные или назальные звуки. Различительная способность спектра, усреднённого на длительном интервале времени для произвольного контекста, ниже, чем сравнение спектров для конкретных гласных. Однако при этом необходимо предварительно распознать тип гласного, а это, в свою очередь, зависит от пола диктора. Формантные частоты спектра, которые обычно отождествляются с резонансными частотами речевого тракта, теоретически менее подвержены влиянию помех и искажений в канале связи. Вместе с тем, автоматическое определение формантных частот сталкивается с серьёзными трудностями, и ни один из известных алгоритмов не гарантирует отсутствие ошибок, включая пропуск или появление ложных оценок. Поэтому в настоящей работе особое внимание уделяется формантному анализу.

Распознавание пола по долговременному спектру вносит задержку в принятие решения, которая далеко не всегда приемлема в технических приложениях. Цель данной работы состоит в определении возможности распознавания пола диктора на сравнительно коротком сегменте речи по физиологически адекватным признакам: параметрам модели голосового источника и формантным частотам на сегменте гласного звука, выделенного из слова.

Речевой материал

Как и в [1], в настоящей работе использовалась база данных числительных русского языка от 0 до 9 для 49 мужчин и 37 женщин, определённых путём кластеризации в пространстве формантных частот как характерные представители из множества голосов 243 мужчин и 186 женщин. Каждый диктор произнёс от 400 до 800 слов через 4 типа микрофонов, расположенных в ближнем поле на разном расстоянии от диктора. В экспериментах использовалось 26404 произнесения слов мужчинами и 15109 — женщинами.

Распознавание пола диктора выполнялось на сегментах ударных гласных длительностью от 70 до 200 мс. Эти сегменты были найдены специальным алгоритмом, описанным в [17].

Метод классификации

В задаче автоматического распознавания диктора функция плотности вероятности в пространстве параметров обычно аппроксимируется с помощью взвешенной суммы нормальных распределений (так называемая Gauss Mixture Model — GMM). Трудность использования такого подхода состоит в неопределённости выбора числа компонент и начальных значений величин математических ожиданий и дисперсии. В настоящей работе распознавание пола диктора выполнялось с использованием модифицированного метода Парзена для аппроксимации функции многомерной плотности вероятности в пространстве параметров. Метод Парзена относится к классу ядерных непараметрических оценок плотности распределения вероятности [18]. Этот метод использует физически правдоподобное предположение, что в окрестности каждого вектора, принадлежащего некоторому классу объектов, находятся вектора, также принадлежащие этому классу, причём вероятность появления этих векторов убывает по мере удаления от исходного

вектора. Естественно принять закон убывания в виде нормального распределения.

В многомерном случае с гауссовым ядром плотность вероятности $P(X)$ представляется как

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

где X_i — независимые и одинаково распределённые наблюдения некоторой случайной величины, N — размер выборки, σ — среднеквадратичное отклонение, являющееся функцией от числа наблюдений N, $\sum$ — ковариационная матрица. Поскольку плотность вероятности восстанавливается в малой окрестности каждого вектора из обучающей выборки, можно принять, что ковариационная матрица $\sum$ диагональная.

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

Заметим, что аппроксимация по Парзену с гауссовыми ядрами является предельным случаем метода GMM, когда математическое ожидание каждого элемента смеси помещается в каждый вектор обучающей выборки.

Оценка параметров голосового источника

Технология оценки параметров модели голосового источника, принятая в данной работе, заключается в следующем. На первом этапе выполняется обратная фильтрация с формированием сигнала-остатка R, предположительно содержащего в себе сигнал голосового источника. На втором этапе методом поиска условного минимума среднеквадратической ошибки между интегрированным сигнал-остатком R и моделью голосового источника D [19] вычисляется производная от объёмной скорости потока через голосовую щель $w'(t)$. Поскольку в этой задаче нет гарантии нахождения глобального минимума ошибки, то выбирается решение, которое даёт наименьшую ошибку при запуске процесса минимизации с различными начальными условиями. Метод формирования кодовой книги для начальных условий описан в [20].

На третьем этапе $w'(t)$ интегрируется с тем, чтобы получить объёмную скорость $w(t)$. Для того чтобы исключить появление постоянной составляющей после интегрирования, на модель накладывается условие равенства площадей под положительной и отрицательной ветвью сигнала D:

$$\int_{0}^{T_0} w'(t)dt = 0$$

Это уменьшает число параметров в модели D с 5 до 4.

На четвёртом этапе для вычисления функции площади голосовой щели $\tilde{S}(t)$ используется уравнение потока через голосовую щель:

$$\rho_0 h w' + k_v h w + \frac{c_x \rho_0}{2\tilde{S}} w^2 = \Delta p \tilde{S}\,, \qquad (1)$$

где w — объёмная скорость потока через голосовую щель, ρ_0 — плотность воздуха, h — глубина голосовой щели вдоль оси потока (≈ 0,5 см), k_v — коэффициент вязкого трения для капиллярного канала, Δp — перепад давления над голосовой щелью (≈1500 Па), c_x — коэффициент динамического сопротивления, зависящий от формы голосовой щели и числа Рейнольдса. Все эти параметры принимаются постоянными и определяются заранее [21]. Относительно $\tilde{S}(t)$ (1) является алгебраическим уравнением второго порядка. Поскольку вычисленная по речевому сигналу объёмная скорость w представлена в произвольных единицах, то $\tilde{S}(t)$ нормируется к максимальному значению $\tilde{S}_{\max}$ = 0,2 см².

Затем определяются параметры модели функции площади голосовой щели $S(t)$ путём поиска условного минимума среднеквадратической ошибки между $\tilde{S}(t)$ и $S(t)$. Модель функции площади голосовой щели является обобщением модели [22]:

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

Здесь S_{max} — максимальная площадь открытия голосовой щели (S_{max}=0,2 см²), T_0 — период основного тона, t_1T_0 — момент максимального открытия голосовой щели, t_2T_0 — момент закрытия голосовой щели, p и q — коэффициенты, определяющие скорость раскрытия и закрытия голосовой щели.

Переход от модели производной от объёмной скорости к модели функции площади голосовой щели обусловлен необходимостью применения физиологически адекватных ограничений при поиске условного минимума. Эти ограничения значительно проще установить для функции площади голосовой щели, чем для производной от объёмной скорости.

В экспериментах по распознаванию пола диктора коэффициенты p и q оказались малоинформативными. Вместо этих параметров оценивалась различительная способность следующих шести параметров, вычисленных на функции $S(t)$:

1) T_0 — период основного тона;

2) t_1 = Argmax($S(t)$)/T_0 — относительный момент времени для максимального значения площади;

3) t_2 = Arg($S(t)$ = 0)/T_0, t>0 — относительный момент времени закрытия голосовой щели;

4) Δt_1 = (Argmax($S(t)$)/T_0, t>0) — (Argmax($S'(t)$>0)/T_0) — интервал между моментом времени, в котором площадь голосовой щели принимает максимальное значение, и моментом времени, в котором производная от $S(t)$ максимальна;

5) Δt_2 = (Arg($S(t)$ = 0)/T_0, t>0) — (Argmin($S'(t)$<0)/T_0) — интервал между моментом времени, в котором площадь голосовой щели принимает нулевое значение, и моментом времени, в котором производная от $S(t)$ принимает минимальное значение;

6) $S'_{min}(t)$ — минимальное значение производной от площади.

Использование относительных значений временных параметров позволяет в значительной степени ликвидировать их зависимость от периода основного тона.

На рис. 1–3 показаны трёхмерные распределения ряда этих параметров для объединённых данных мужчин и женщин.

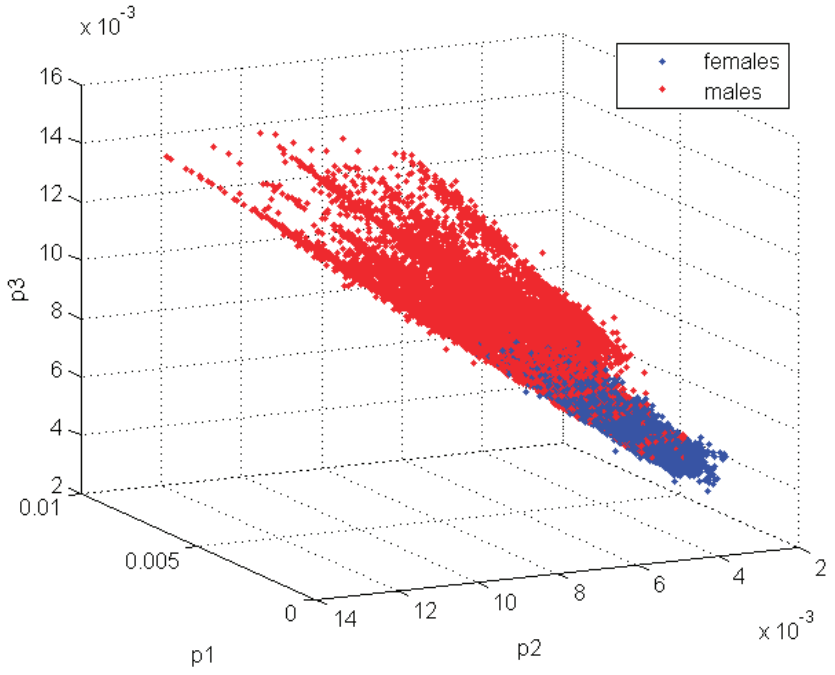


Рис. 1. Распределение парметров T_0, t_1 и t_2 (разметка осей в сек)

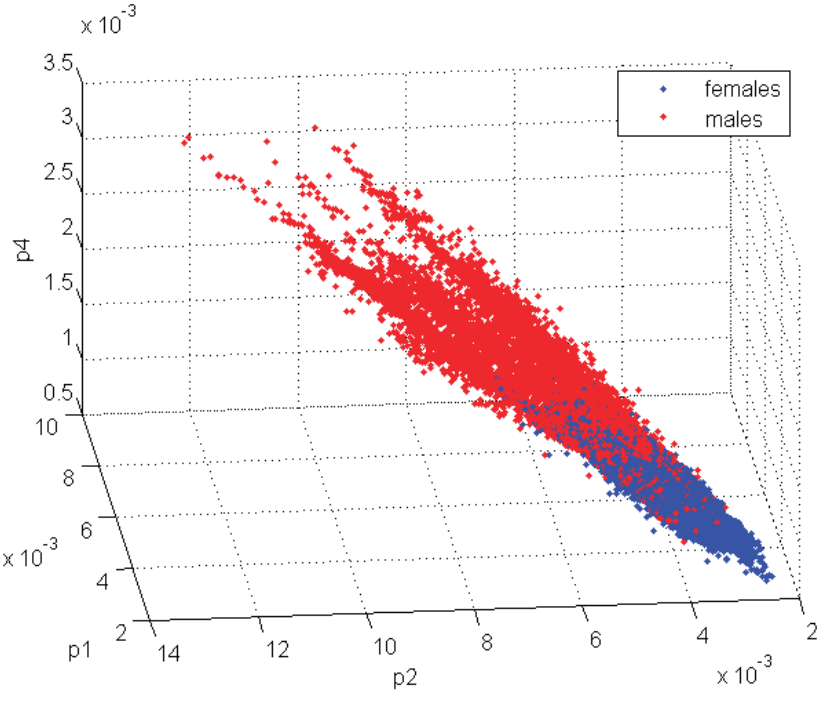


Рис. 2 . Распределение параметров t_1, t_2 и Δt_1

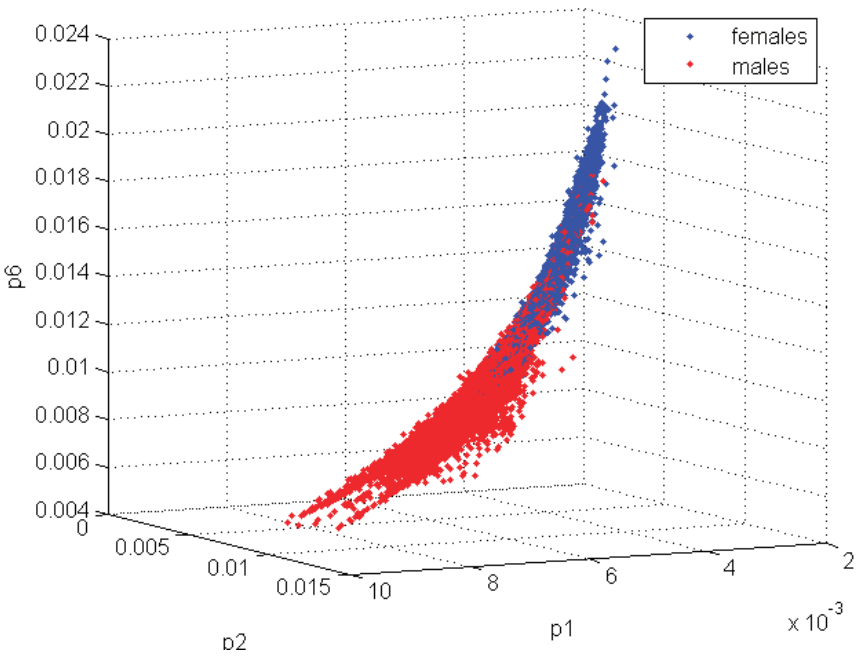


Рис. 3. Распределение параметров t_1, t_2 и $S'_{min}(t)$

Каждый из параметров отдельно демонстрирует весьма высокий уровень суммарной ошибки, т.е. вероятность принять мужской голос за женский, и наоборот (табл. 1).

Таблица 1

Суммарная ошибка распознавания пола по отдельным параметрам

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

Как упоминалось выше, в исследованиях по непосредственному измерению воздушного потока через голосовую щель [3, 4] было найдено, что отношение длительности интервала открытой голосовой щели к периоду основного тона у женщин больше, чем у мужчин. На рис. 4 показана доля (%) интервалов открытой голосовой щели t_2, найденных в мужских и женских голосах в результате определения параметров голосового источника.

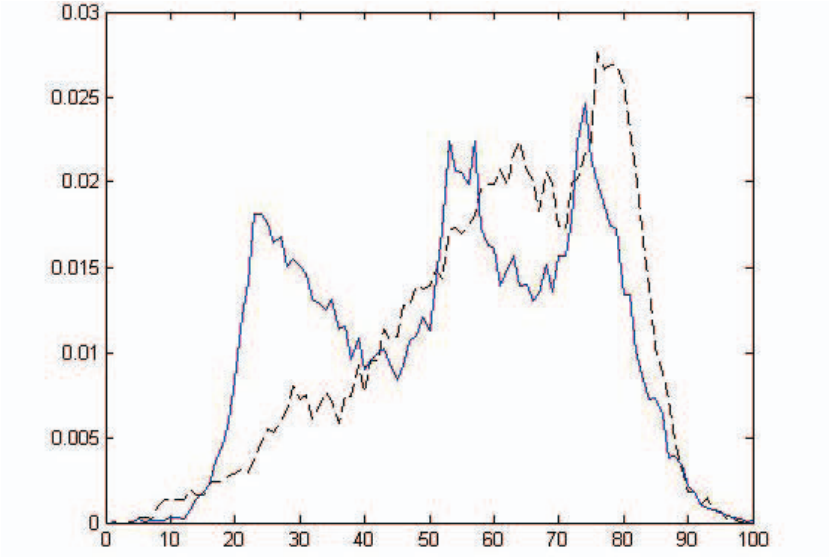


Рис. 4. Распределение относительной величины интервала открытой голосовой щели t_2 для мужских (—) и женских голосов (---) (по оси абсцисс — t_2, %)

Эти данные в целом подтверждают мнение о более длинном интервале открытой голосовой щели у женщин по сравнению с мужчинами, но это различие не слишком велико. Обращает на себя внимание и то, что мужские голоса по этому параметру группируются в три кластера. Нужны дальнейшие исследования для того, чтобы определить, насколько объективна такая группировка.

Как видно из таблицы 1, 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 её закрытия.

Ошибка распознавания пола в двумерном 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 трёхмерных пространствах параметров ошибки распознавания пола значительно ниже, чем в одномерных и двумерных, причём в четырёх пространствах ошибка оказалась ниже 3%, а в двух — ниже 2% (табл. 2).

Таблица 2

Погрешность распознавания пола в трёхмерных пространствах

Параметры	Вероятность ошибки, (%)
t_1, t_2, T_0	2,71
t_1, t_2, Δt_1	2,76
t_1, t_2, $S'_{min}(t)$	2,97
t_1, T_0, Δt_1	1,58
t_1, T_0, $S'_{min}(t)$	1,80
t_2, T_0, Δt_1	2,37

По сравнению с работой [1], в которой была получена суммарная ошибка около 10% на той же базе данных, в настоящей работе ошибки снижены почти в 6 раз. Такое снижение ошибки распознавания было достигнуто как за счёт 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 распознавание пола диктора на относительно коротком вокализованном сегменте речевого сигнала.

Вместе с тем, необходимо принимать во внимание, что обратная задача относительно параметров голосового источника является некорректной. В [23] показано, что эта задача имеет единственное, но неустойчивое решение, когда точно известны резонансные частоты речевого тракта. Однако они могут быть определены однозначно, только когда известен источник возбуждения [24]. Но и в этом случае решение неустойчиво относительно возмущения входных данных. Неоднозначное и неустойчивое решение задачи определения параметров голосового источника — её принципиальное свойство.

На одном из речевых сегментов может произойти отказ от анализа параметров голосового источника в силу нестабильности его оценки с использованием метода обратной фильтрации. Решение о поле диктора будет задержано до тех пор, пока не появится сегмент с хорошим качеством анализа. Поэтому для более надёжного распознавания пола целесообразно дополнительно привлечь информацию о распределении формантных частот.

Формантный анализ

Резонансные частоты речевого тракта определяются его размерами и формой, и, следовательно, несут информацию об анатомии тракта и поле человека. Однако, как упоминалось выше, задача определения резонансных частот тракта по речевому сигналу является некорректной, т.е. не имеет единственного и устойчивого решения. При разных подходах к определению резонансных частот тракта их отождествляют с частотами локальных пиков спектра мощности (формант), полюсами передаточной функции тракта, найденной методом линейного предсказания, мгновенными частотами некоторой функции и т.д. Каждый их этих методов страдает склонностью к обнаружению лишних (ложных) резонансов и потере истинных [25]. Это является следствием принципиальной некорректности обратной задачи относительно резонансных частот речевого тракта.

Вместе с тем, эту обратную задачу можно приближённо решить, используя дополнительную информацию о процессах речеобразования и ограничения на совместное распределение резонансных частот в данном языке. Один из наиболее важных факторов заключается в различии между свойствами речевого сигнала на интервале открытой и закрытой голосовой щели. Во время открытой голосовой щели частотная характеристика речевого тракта искажается влиянием источника голосового возбуждения, тогда как при закрытой щели предполагается существование свободных затухающих колебаний с частотами, которые определяются только резонансами тракта. Поэтому и анализ резонансных частот целесообразно выполнять на интервале закрытой голосовой щели.

Определение интервала времени, на котором голосовая щель закрыта — сложная задача, учитывая, что у женских голосов этот интервал может вообще отсутствовать. Поэтому точность определения этого интервала невелика, и можно говорить лишь о приближённом определении его положения. Обычно этот интервал определяется относительно пика сигнала-остатка, найденного методом линейного предсказания. И хотя этот пик достаточно хорошо детектируется этим методом во временной области, положение и длительность интервала закрытой щели остаются весьма неопределёнными.

Альтернатива временному анализу состоит в использовании спектрально-временных характеристик речевого сигнала. Хорошо известно, что квазипериодическая последовательность импульсов голосового возбуждения видна на сонограммах речевого сигнала в виде почти вертикальных линий (рис. 5).

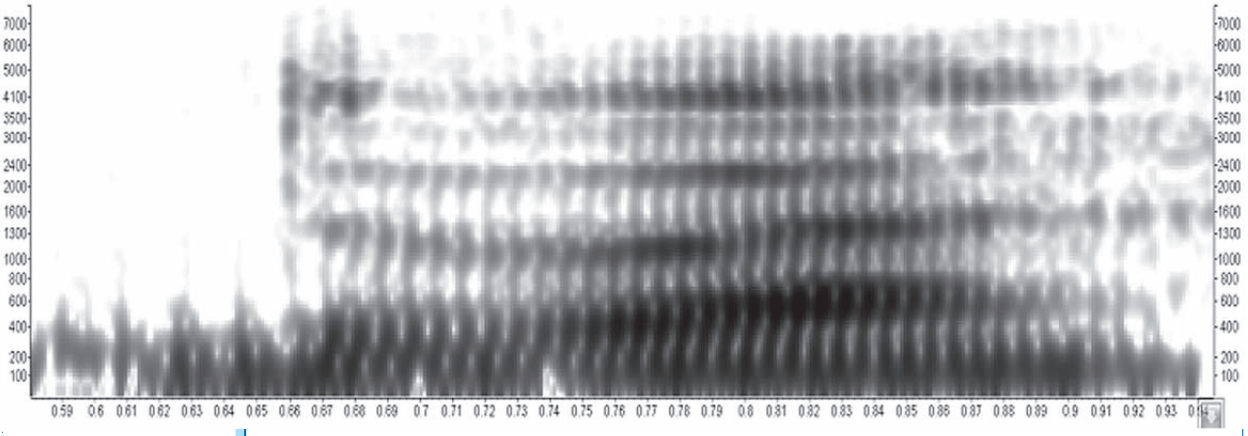


Рис. 5. Сонограмма слова «*два*» (по оси ординат — *мел-частоты*, по оси абсцисс — секунды)

Суммируя в каждый момент времени амплитуды в каждой частотной полосе, можно получить временную последовательность изменения энергии речевого сигнала во времени (рис. 6). Речевой сигнал (рис. 6а) был синтезирован с участием пяти формант при использовании модели площади голосовой щели (2) и голосового источника, вычисленного по [26]. На осциллограмме речевого сигнала вертикальными линиями отмечены моменты закрытия и открытия голосовой щели, а между ними находится маркер максимума возбуждения, найденный путём обратной фильтрации.

Сопоставляя последовательность всплесков энергии огибающей речевого сигнала с последовательностью синтетических импульсов голосового возбуждения, можно примерно установить моменты открытия и закрытия голосовой щели. Момент открытия находится в области максимума производной огибающей речевого сигнала, а момент её закрытия — в окрестности минимума огибающей. Это правило работает и для женских голосов с высоким основным тоном, только в этом случае интервал между максимумом и минимумом производной, скорее всего, следует интерпретировать как интервал с наибольшим раскрытием голосовой щели.

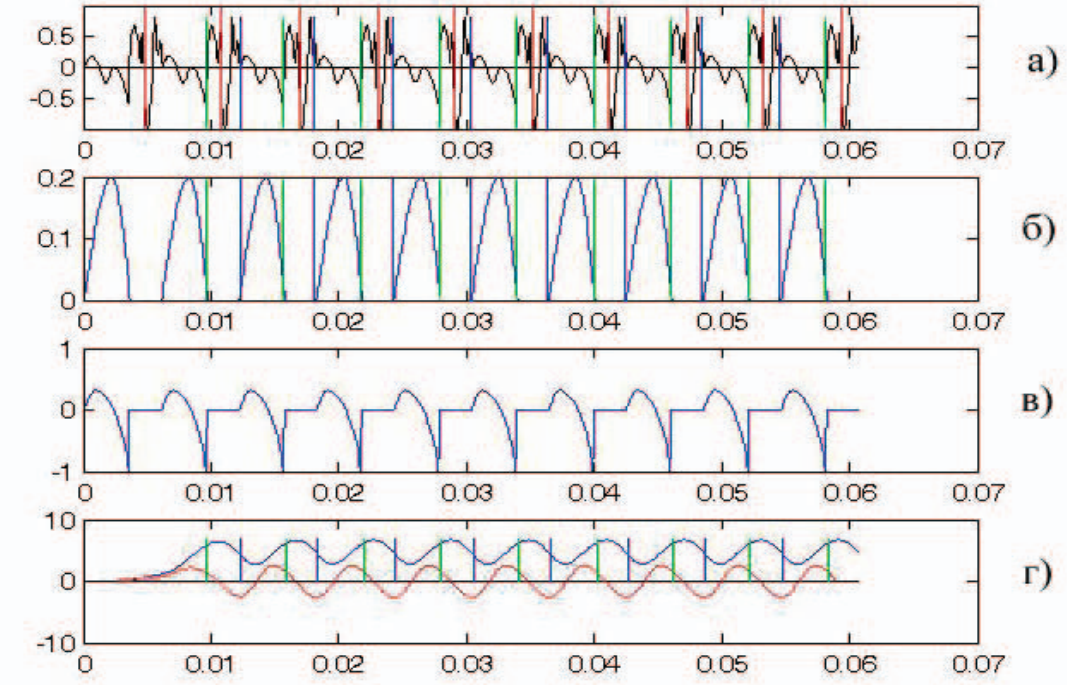


Рис. 6. Синтезированный речевой сигнал (а), площадь голосовой щели (б), импульсы голосового возбуждения (в), огибающая энергии сигнала во времени (г, вверху) и её производная (г, внизу)

Оценка резонансных частот речевого тракта выполняется в три этапа. На первом этапе определяются частоты полюсов передаточной функции методом линейного предсказания на интервалах закрытой голосовой щели на некотором вокализованном сегменте речевого сигнала. Этим частотам приписываются значения частот медиан гистограммы найденных мгновенных оценок $\{F_{1л}, F_{2л}, F_{3л}, \ldots\}$. На втором этапе на этом же сегменте речи находится средний спектр, и определяются частоты его локальных максимумов $\{F_{1с}, F_{2с}, F_{3с}, \ldots\}$. Затем сравниваются частоты, найденные этими методами. Если какая-то пара частот расположена достаточно близко (например, на расстоянии, не превышающем 20% от их среднего значения), в качестве кандидата на соответствующую частоту форманты принимается среднее значение для этой пары частот. Если в диапазоне частот первой форманты какой-то метод не получил оценки частоты, то за F_1 принимается значение, найденное другим методом. Если же оба метода не дают правдоподобных оценок $\tilde{F}_1$, то на этом периоде основного тона происходит отказ от формантного анализа. Такой же отказ происходит и в других случаях существенного расхождения оценок, если не удаётся найти оценку, достаточно близкую по вероятности к распределению соответствующей форманты для данного языка. Применяются также и другие способы коррекции оценок формантных частот, которые здесь не приводятся в силу их сложной логики.

На последнем этапе формантного анализа для каждой найденной частоты $\tilde{F}_i$ формируется полосовой фильтр с центральной частотой $\tilde{F}_i$ и шириной полосы пропускания около 20% от $\tilde{F}_i$. К речевому сигналу, пропущенному через этот фильтр, применяется метод оценки средней мгновенной частоты на интервале закрытой голосовой щели, описанный в [27]. Полученные таким образом оценки форматных частот $\{F_1, F_2, F_3\}$ используются для распознавания пола диктора.

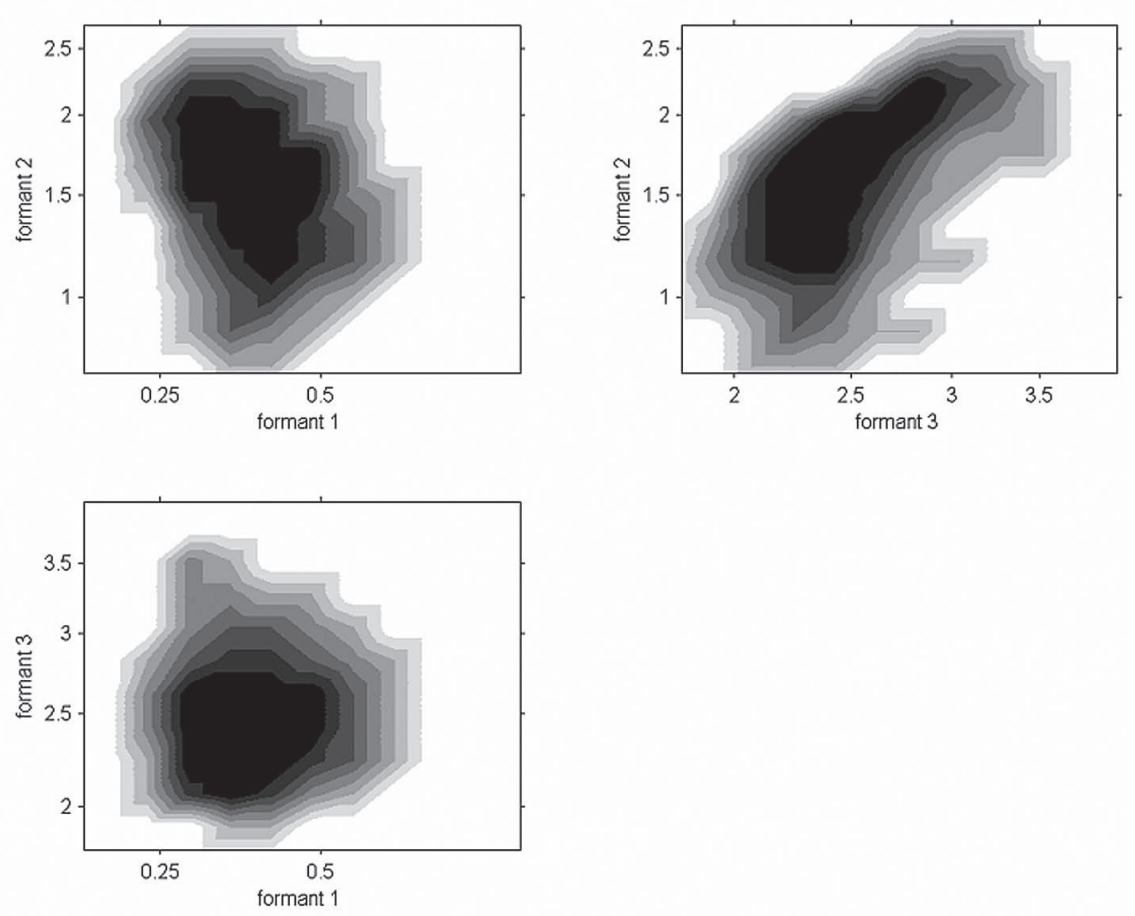
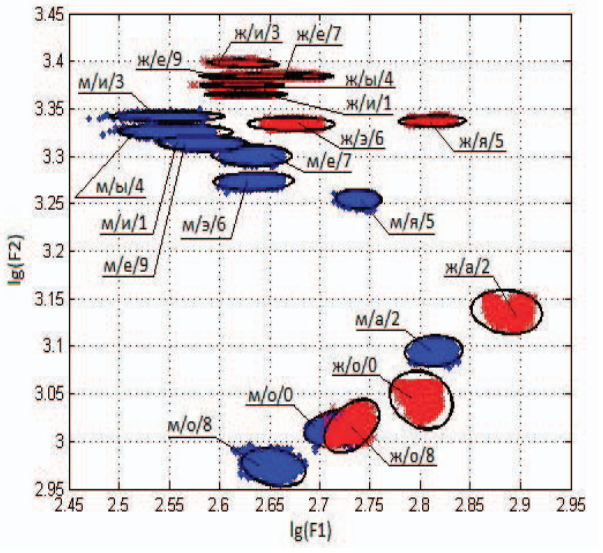


Рис. 7. Проекции распределения первых трёх формантных частот

Анализ совместного распределения формантных частот у мужских и женских голосов для множества вокализованных сегментов, полученных по используемой базе данных, оставляет мало надежд для распознавания пола диктора в таком совместном распределении. На рис. 7 показаны проекции этого распределения на плоскости формантных частот (F_1, F_2), (F_2, F_3) и (F_1, F_3). Как видно, наиболее вероятные значения формантных частот тяготеют к середине частотного диапазона.

Иная картина обнаруживается при рассмотрении частотных распределений порознь для каждого ударного гласного в исследованных словах. На рис. 8 показаны распределения первых трёх формантных частот, вычисленных на ударных гласных числительных русского языка от 0 до 9. На этом рисунке в подписях первый символ *м* или *ж* соответствует полу диктора. В косых скобках указан тип гласного, последний символ соответствует числительному, в котором этот гласный находится. Вычисленные формантные частоты квантовались с шагом 10% для первой форманты и 15% — для второй и третьей.



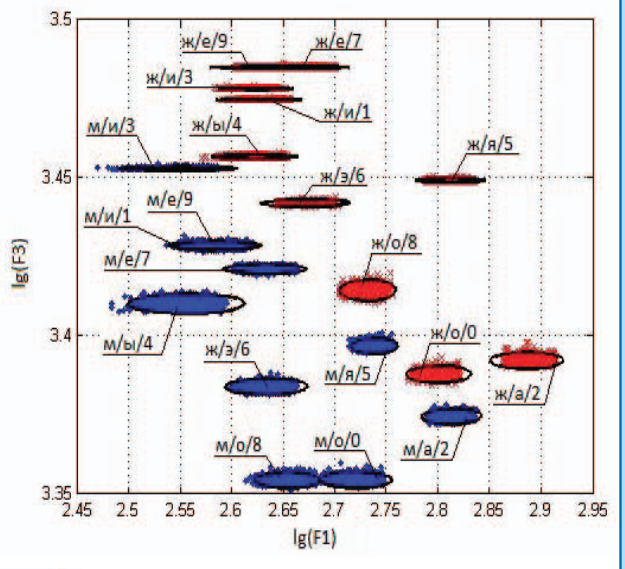
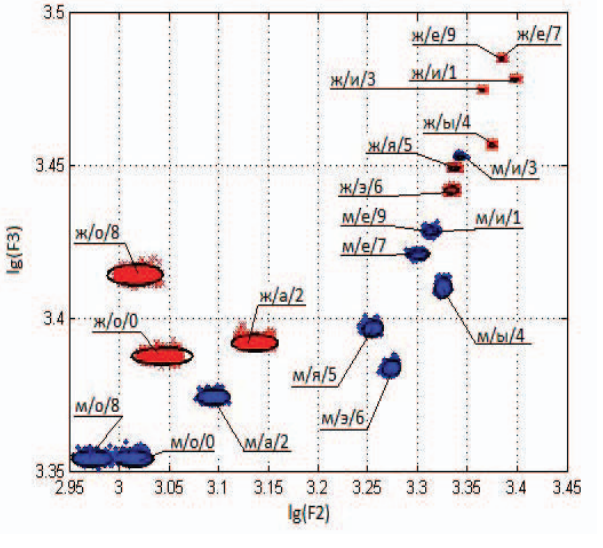


Рис. 8. Распределение формантных частот для ударных гласных в числительных от 0 до 9



Видно, что распределение формантных частот на плоскости (F_1, F_2) для гласного /*о*/ в слове «*восемь*» женского голоса практически пересекается с распределением формантных частот гласного /*о*/ в слове «*ноль*» мужского голоса, тогда как формантные частоты гласного /*о*/ в слове «*ноль*» женского голоса далеко разнесены от формантных частот в слове «*восемь*» мужского голоса. Эти же слова разделены на плоскостях (F_1, F_3) и (F_2, F_3).

В таблицах 3–5 приведены ошибки распознавания пола в одномерных, двумерных и трёхмерном пространствах распределения формантных частот.

Таблица 3

Ошибка распознавания пола (%) по отдельным формантным частотам гласных звуков

Гласный	/о/	/и/	/а/	/ы/	/я/	/э/	/е/	Среднее
F_1	23	2	0	0,5	0	6	46	11
F_2	40	0	0	0	0	0	0	5
F_3	0	0	0	0	0	0	0	0

Таблица 4

Ошибка распознавания пола (%) по парам формантных частот

Гласный	/о/	/и/	/а/	/ы/	/я/	/э/	/е/	Среднее
(F_1, F_2)	22	0	0	0	0	0	0	3
(F_1, F_3)	0	0	0	0	0	0	0	0
(F_1, F_3)	0	0	0	0	0	0	0	0

Таблица 5

Ошибка распознавания пола (%) в трёхмерном пространстве

Гласный	/о/	/и/	/а/	/ы/	/я/	/э/	/е/	Среднее
(F_1, F_2, F_3)	0	0	0	0	0	0	0	0

Распределение формантных частот для каждого гласного может быть описано своей функцией. Это даёт возможность распознавания пола без предварительного распознавания типа гласного. Если измеренный вектор формантных частот попадает в область распределения, в которой отношение правдоподобия близко 0,5, то не имеет смысла выполнять распознавание и надёжнее отказаться от принятия решения. Во всех остальных случаях можно определить отношение правдоподобия путём вычисления вероятности принадлежности измеренного вектора частот к мужским или женским голосам для распределений в данной частотной области.

Зависимость параметров

Три наилучших параметра модели голосового источника и три формантные частоты составляют вместе шестимерное пространство, в котором необходимо принимать решение о поле диктора. Поскольку объём обучающей выборки весьма ограничен, то следует ожидать, что доверительный интервал для оценки вероятности распознавания окажется неприемлемо большим. Если же зависимость между параметрами голосового источника и форман-

тными частотами окажется мала, то целесообразно оценивать правдоподобие принадлежности к некоторому полу раздельно в каждом из этих подпространств. Как видно из таблиц 6–7, корреляция между этими подпространствами действительно очень мала даже для пары (F_1, T_0).

Таблица 6

Коэффициент корреляции между формантными частотами и параметрами модели голосового источника (мужчины)

Форманта	t_1/T_0	t_2/T_0	T_0
F_1	0,0189	0,0300	0,0532
F_2	-0,0188	-0,0165	0,0126
F_3	-0,0110	-0,0265	0,0208

Таблица 7

Коэффициент корреляции между формантными частотами и параметрами модели голосового источника (женщины)

Форманта	t_1/T_0	t_2/T_0	T_0
F_1	0,0587	0,0125	0,0358
F_2	-0,0429	0,0024	-0,0181
F_3	-0,0359	0,0003	-0,0144

Считается, что существует положительная корреляция между частотой основного тона F_0 и частотой первой форманты F_1. Это представление исходит из соображений о связи анатомических размеров тела. Действительно, и частота основного тона, и частота формант у женщин обычно выше, чем у мужчин, хотя нередко встречаются и обратные примеры. Вместе с тем, анализ зависимости F_1 от F_0 выявил неожиданную и сложную картину их взаимоотношений. На рис. 9 показаны распределения на плоскости (F_0, F_1) и (F_0, F_2), полученные по всем ударным гласным для мужчин (синий цвет) и женщин (красный цвет). На рис. 10 эти распределения показаны отдельно для мужских и женских голосов.

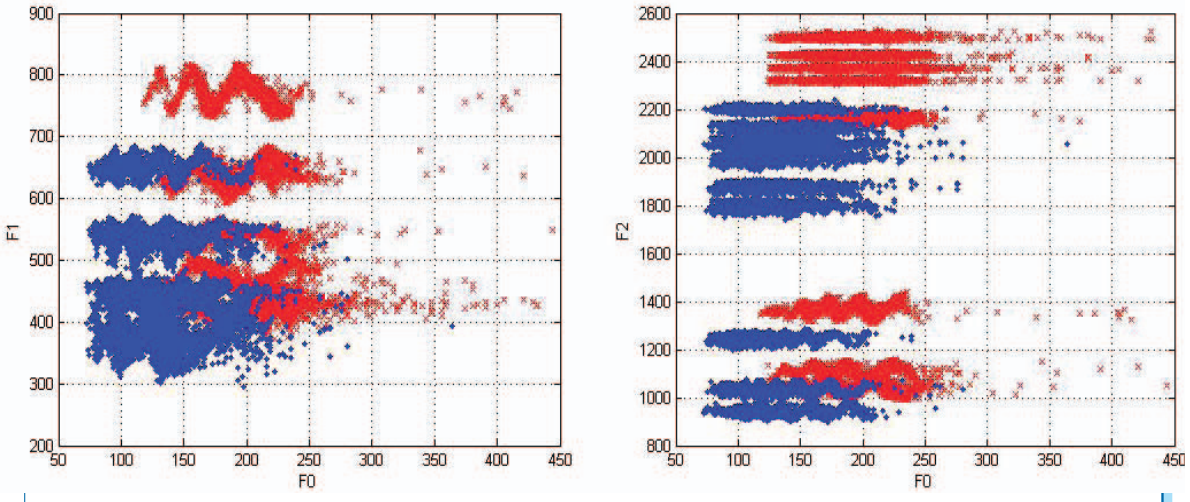


Рис. 9. Распределения (F_0, F_1) — слева; (F_0, F_2) — справа

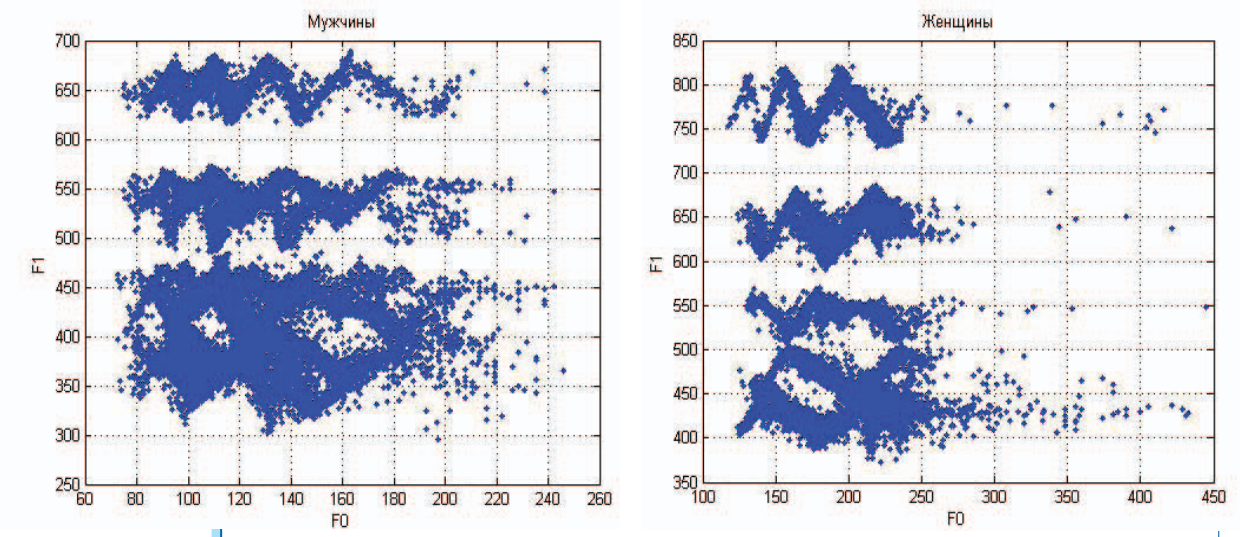


Рис. 10. Распределения (F_0, F_1): мужские голоса — слева, женские — справа

Из этих рисунков видно, что наблюдаются как положительные, так и отрицательные зависимости между этими параметрами — в среднем корреляция оказывается малой. Особенно ясно проявляется чередование положительной и отрицательной зависимостей между частотой основного тона и частотой первой форманты для каждого отдельного гласного (рис. 11). Механизмы найденных зависимостей измеряемой формантной частоты от частоты основного тона требуют специального исследования.

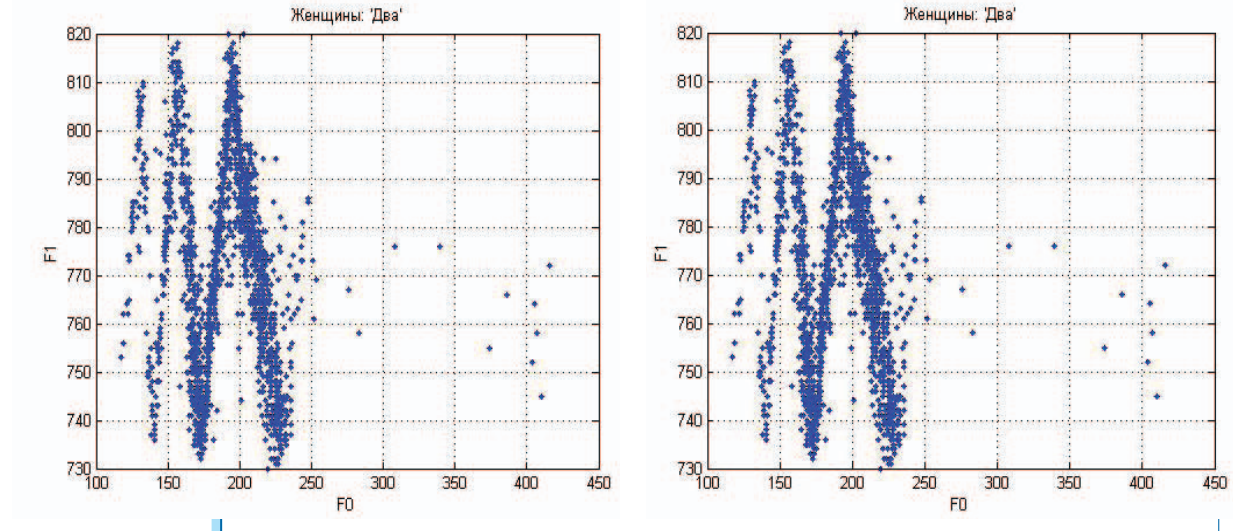


Рис. 11. Распределения (F_0, F_1) для ударной гласной */а/* в слове «*два*» по всем мужским голосам — слева и женским — справа

Совместное решение

Если на некотором речевом сегменте получены решения и о параметрах голосового источника, и о формантных частотах, достигается минимальная вероятность ошибки распознавания с достаточно хорошим запасом надёжности.

Поскольку в настоящей работе установлена малая зависимость параметров голосового источника и формантных частот, то апостериорная вероятность ошибки распознавания пола на таком сегменте может быть вычислена методом правдоподобия по Байесу для независимых величин. Однако такое решение может быть получено далеко не для всех речевых сегментов, так как информация может быть только о голосовом источнике или только о формантных частотах.

Анализ параметров голосового источника необходимо выполнять на интервале гласно-подобных звуков, поскольку полоса частот на сегментах звонкой и назальной смычек ограничена низкими частотами. Вследствие этого форма голосового источника определяется очень грубо, сохраняя, в основном, лишь период импульсов возбуждения. Аналогично оценка формантных частот так же может выполняться только на сегментах гласно-подобных звуков, точнее, на ударных гласных. Поэтому информация о поле диктора поступает в блок принятия решений не непрерывно, а в соответствии с последовательностью появления гласных в потоке речи. Такое свойство диктует необходимость предварительной сегментации речевого потока на участки, пригодные для анализа параметров голосового источника и формантных частот.

Необходимо также принимать во внимание вероятность того, что на сегменте гласного появится отказ от анализа голосового источника, или определения формантных частот, или того и другого одновременно. Для процедуры анализа голосового источника этот отказ может возникнуть вследствие неустойчивости обратной фильтрации или естественного перекрытия областей существования параметров мужских и женских голосов. Вероятность такого отказа может доходить до 30% от числа анализируемых гласных.

Отказ от результатов формантного анализа также может возникнуть по двум причинам. Во-первых, обучение на тип пола выполняется только на ударных гласных, то на других гласно-подобных сегментах векторы формантных частот могут не попасть в области существования формант, определённых в процессе обучения. Во-вторых, как известно, не существует абсолютно надёжного алгоритма вычисления формантных частот, и на некоторых ударных гласных не все форманты могут быть определены или появятся «лишние» оценки формантных частот.

В результате на сегментах гласных звуков может вообще отсутствовать информация о поле диктора из-за одновременного отказа как анализа параметров голосового источника, так и оценки формантных частот. На таких сегментах всё же может быть получена информация о поле диктора путём измерения периода основного тона T_0, отношения амплитуд первой и второй гармоник в спектре, а также отношения энергии шумовой и гармонической компоненты на периоде основного тона. Различительная способность этих признаков не очень велика, но они более устойчивы. При этом период основного тона может быть найден не только на гласных звуках, но и на звонких и на назальных смычках.

Процесс принятия решения о поле диктора должен включать в себя непрерывную оценку во времени вероятности мужского или женского голоса, доставляемую каждым из методов анализа параметров речевого сигнала. В тот момент, когда объединённая по всем методам оценка превышает заданный порог, принимается окончательное решение. Длительность речевого сегмента к этому моменту может варьироваться от нескольких десятков миллисекунд до секунд, в зависимости от фонетического состава речевого сигнала. Принятие решения на основе частных решений разных типов классификаторов с помощью так называемого агрегирования допускает отсутствие данных на каком-либо сегменте и обычно обеспечивает ошибку распознавания меньше наименьшей ошибки каждого классификатора [28].

Заключение

Распознавание пола диктора, основанное на анализе параметров голосового источника и формантных частот, требует предварительной сегментации речевого сигнала на участки с голосовым возбуждением. Установлено, что параметры голосового источ-

ника и формантные частоты практически независимы, что позволяет применять достаточно простые решающие правила. На тех сегментах речевого сигнала, где получено решение от хотя бы одного метода анализа, ошибка распознания пола ниже 2%, а при совместном решении эта ошибка близка к нулю. Вместе с тем, на некоторых речевых сегментах такое решение может отсутствовать, что приводит к увеличению длительности анализируемого речевого сигнала.

Список литературы

1. *Сорокин В.Н., Макаров И.С*. Распознавание пола диктора по голосу // Акустический ж., 2008. Т. 54. № 4. С.1— 9.

2. *Holmberg E.B., Hillman R.E., Perkell J.S*. Glottal airflow and transglottal air measurements for male and female speakers in soft, normal, and loud voice // J. Acoust. Soc. Amer., 1988. V 84. № 2. P. 511–529.

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

4. *Holmberg E.B., Hillman R.E., Perkell J.S., Guiod P., Goldman S.L*. Comparison among aerodynamic, electroglottographic, and acoustic spectral measures of female voice // J. Speech Hear. Res., 1995. V. 38. P. 511–529.

5. *Price P.J*. Male and female voice source characteristics: inverse filtering results // Speech Communication., 1989. V. 8. P. 261–277.

6. *Wu Ke, Childers D.G*. Gender recognition from speech. Part 1: Coarse analysis // J. Acoust. Soc. Amer., 1991. V. 90. 4. Pt.1, P. 1828–1840.

7. *Childers D.G., Wu Ke*. Gender recognition from speech. Part 2: Fine analysis // J. Acoust. Soc. Amer.. 1991. V. 90. № 4. Pt.1, P. 1841–1856.

8. *Hanson H.M*. Glottal characteristics of female speakers: Acoustic correlates // J. Acoust. Soc. Amer., 1997. V. 101. P. 466–481.

9. *Hanson H.M., Chuang E.S*. Glottal characteristics of male speakers: Acoustic correlates and comparison with female data // J. Acoust. Soc. Amer., 1999. V. 106. № 2, P. 1064–1077.

10. *Iseli M., Shue Y.-L., Alwan A*. Age, sex, and vowel dependencies of acoustic measures related to the voice source // J. Acoust. Soc. Amer., 2007. V. 121. P. 2283–2295.

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

12. *Sigmund M*. Gender distinction using short segments of speech signal. Int. // J. of Computer Science and Network Security, 2008. V. 8. № 10. P. 159–163.

13. *Hillenbrand J., Getty L.A., Clark M.J., Wheeler K*. Acoustic characteristics of American English vowels // J. Acoust. Soc. Amer., 1998. V. 97. P. 3099–3111.

14. *Whiteside S.P*. Identification of a speaker's sex: a study of vowels // *Percept. Mot. Skills*, 1998. V. 8. № 2. P. 579–584.

15. *Lavner Y., Gath I., Rosenhouse J*. The effects of acoustic modifications on the identification of familiar voices speaking isolated vowels // Speech Communication, 2000. V. 30. P. 9–26.

16. *Smith D.R, Walters T.C, Patterson R.D*. Discrimination of speaker sex and size when glottal-pulse rate and vocal-tract length are controlled // J Acoust Soc Am., 2007. V. 122. № 6. P. 3628–3639.

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

18. *Parzen E*. An estimation of a probability density function and mode // Ann.Math.Stat., 1962. V. 33, P. 1065–1076.

19. *Ananthapadmanabha T*. Acoustic Analysis of Voice Source Dynamics // STL-QPSR, 1984. № 2–3. P. 1–24.

20. *Сорокин В.Н., Тананыкин А.А*. Начальные условия в задаче верификации голосового источника // Информационные процессы, 2010. Т. 10. № 1. С. 1–10.

21. *Сорокин В.Н*. Теория речеобразования. Радио и связь, 1985.

22. *Lin Q*. Nonlinear Interaction in Voice Production // STL-QPSR, 1987. № 1. P. 1–12.

23. *Леонов А.С., Сорокин В.Н*. О единственности определения голосового источника по речевому сигналу и формантным частотам // Доклады Академии Наук, 2012. Т. 444. № 5. С. 492–495.

24. *Леонов А.С., Сорокин В.Н*. Об однозначности определения резонансных частот голосового тракта по речевому сигналу // Доклады Академии наук, 2011. Т. 440. № 1. С. 32–34.

25. *Леонов А.С., Макаров И.С., Сорокин В.Н*. Устойчивость оценок формантных частот // Речевые технологии, 2009. № 1. С. 3–18.

26. *Сорокин В.Н*. Синтез речи. М.: Наука, 1992.

27. *Леонов А.С., Макаров И.С., Сорокин В.Н*. Частотные модуляции в речевом сигнале // Акустический ж., 2009. Т. 55. № 6. С. 809–821.

28. *Сорокин В.Н., Вьюгин В.В., Тананыкин А.А*. Распознавание личности по голосу: Аналитический обзор // Информационные процессы, 2012. Т. 12. № 1. С. 1–30.

Сведения об авторах:

Сорокин Виктор Николаевич —
доктор физико-математических наук, ведущий научный сотрудник Института проблем передачи информации РАН. Занимается фундаментальными исследованиями процессов речеобразования и восприятия речи и приложениями к речевым технологиям с 1964 г. Опубликовал более 130 работ, в том числе монографии «Теория речеобразования» (1985, Радио и Связь) и «Синтез речи» (1992, Наука).

Ромашкин Юрий Николаевич —
кандидат технических наук, Окончил Московский инженерно-физический институт, факультет «Автоматика и электроника». Область научных интересов: цифровая обработка речевых сигналов, фильтрация речи на фоне помех, автоматическое распознавание речи и языка, идентификация говорящего по голосу, низкоскоростное кодирование речи, оценка качества трактов речевой связи.
E-mail: romayn@yandex.ru

Тананыкин Александр ААлександрович —
аспирант, ИППИ РАН, окончил МИЭТ (ТУ). Область научных интересов: обработка речи, математическое моделирование, программирование.
E-mail: tananykin@mail.ru