Информационные процессы, Том 13, №2, 2013, стр. 35-47

====== ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ В ТЕХНИЧЕСКИХ ======
====== И СОЦИАЛЬНО-ЭКОНОМИЧЕСКИХ СИСТЕМАХ =======

ОЦЕНКА ДЛИНЫ РЕЧЕВОГО ТРАКТА

В.Н.Сорокин, И.В.Гераськин

Институт проблем передачи информации, Российская академия наук, Москва, Россия
Поступила в редколлегию 25.04.2013

Аннотация. Исследовались два способа оценки длины речевого тракта, эквивалентной длине однородной акустической трубы. Один способ был основан на вычислении длины тракта по разности частот ближайших локальных максимумов спектра, находящихся выше 4 кГц. В другом способе длина тракта вычислялась по средней частоте второй форманты, определенной по частотам первых трех формант. Также рассматривались варианты анализа независимо от контекста и в зависимости от известных гласных звуков.

Вероятность правильного распознавания пола диктора обоими способами составила около 13%, и практически не зависит от знания контекста. Вероятность правильного распознавания мужских голосов по разности высших формант независимо от контекста составила около 31 %, а женских голосов - около 25.5%, а в зависимости от контекста – 37% и 31%, соответственно. При контекстно-независимом распознавании по средней частоте второй форманты вероятность правильного распознавания мужских голосов близка к 27%, а женских – к 21.5%, тогда как при контекстно-зависимом распознавании с известным типом гласного вероятность правильного распознавания мужских голосов повышается до 43 %, а женских - до 35.5%.

Ключевые слова: длина речевого тракта, распознавание пола, распознавание диктора.

1. ВВЕДЕНИЕ

При решении обратной задачи вариационным методом относительно формы речевого тракта используется параметрическая модель тракта, в число параметров которой входит длина речевого тракта L [16]. С целью сокращения объема перебора значений параметров модели желательно заранее оценить наиболее вероятную длину тракта в текущем произнесении. Для каждого диктора не только характерна длина речевого тракта в нейтральном состоянии, но и управление высотой гортани для артикуляции разных звуков может отличаться от других дикторов. Поэтому оценка длины речевого тракта может быть полезной и при распознавании дикторов.

В задаче автоматического распознавания речи независимо от диктора целесообразно предварительно определить пол диктора, поскольку акустические характеристики мужских и женских голосов различаются. Анатомические измерения показывают, что различие в длине речевого тракта между мужчинами и женщинами может доходить до 25% [7]. Идентификация пола диктора еще до начала процесса распознавания может улучшить вероятность правильного распознавания речи. В этих задачах вместо непосредственной оценки длины речевого тракта диктора обычно применяют технологию компенсации различия в длине тракта у разных дикторов. Такая компенсация может выполняться путем трансформации параметров речевого сигнала во временной, частотной или кепстральной области. Во временной области, наряду с растяжением/сжатием оси времени, используется преобразование Меллина [18]. В частотной области применяется линейное или нелинейное преобразование оси частот [1, 8]. Иногда такое преобразование проводится только по усредненной частоте 3-й форманты [5], либо по средним частотам первых трех формант [19]. В кепстральной области применяют вычитание кепстра, усредненного на длительном интервале времени, а также кепстра, вычисленного на предыдущем кадре [13].

Компенсация длины тракта диктора является косвенным методом нормализации анатомических параметров. Непосредственная оценка длины тракта в дополнение к компенсации ее вариаций, может повысить устойчивость систем распознавания речи к индивидуальным особенностям анатомии дикторов.

Для оценки длины речевого тракта использовались модели тракта различной сложности. В [11] с этой целью применялась двух-трубная модель тракта. В [2, 3] предлагается эвристическая оценка $L=(L_1+L_2+L_3)/3$, где

$$L_k=c_0(2k\text{-}1)/4F_k, \quad (1)$$

а c_0 – скорость звука (≈350 м/с), F_k – частота k-го резонанса. Это выражение справедливо только для однородной акустической трубы, но, по сообщению авторов этих работ, полученная оценка коррелирована с истинной длиной речевого тракта. В [4] сообщается о погрешности около 3% при оценке длины речевого тракта с помощью определения параметров скрытой марковской модели, на входе которой используются формантные частоты.

В настоящей работе оценка длины речевого тракта выполнялась двумя способами: через оценки частот высших формант и по оценке частоты второй форманты как средней частоте первых трех формант на интервале гласных звуков.

2. ЧАСТОТЫ ВЫСШИХ РЕЗОНАНСОВ

Длина речевого тракта определяется как расстояние вдоль его средней линии от голосовой щели до последней координаты на губах. В свою очередь, средняя линия есть геометрическое место точек, находящихся посередине кратчайшей прямой линии между любой точкой на неподвижной поверхности речевого тракта и соответствующей точкой на подвижной поверхности. Положение средней линии и ее длина зависят не только от высоты гортани, но и от формы речевого тракта (Рис. 1).

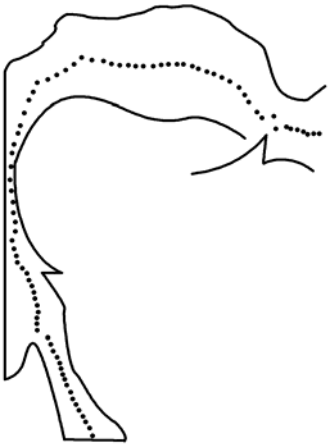


Рис. 1. Форма речевого тракта и его средняя линия для гласного /*a*/ [15].

Эффективная длина тракта несколько увеличивается в силу особенностей излучения звука в пространство. Это увеличение зависит от площади ротового отверстия и может достигать 5% при большом раскрытии рта [14].

В [15] было показано, что для неоднородной акустической трубы резонансная частота увеличивается или уменьшается в зависимости от того, в каком месте тракта происходит сужение. Принимая, для простоты, что возмущение исходной площади поперечного сечения $S(x)$ сконцентрировано в одной точке с координатой x_0, т. е. описывается δ-функцией, $S_1(x)=\varepsilon\delta(x-x_0)$, то поправки на собственные числа возмущений системы находятся как

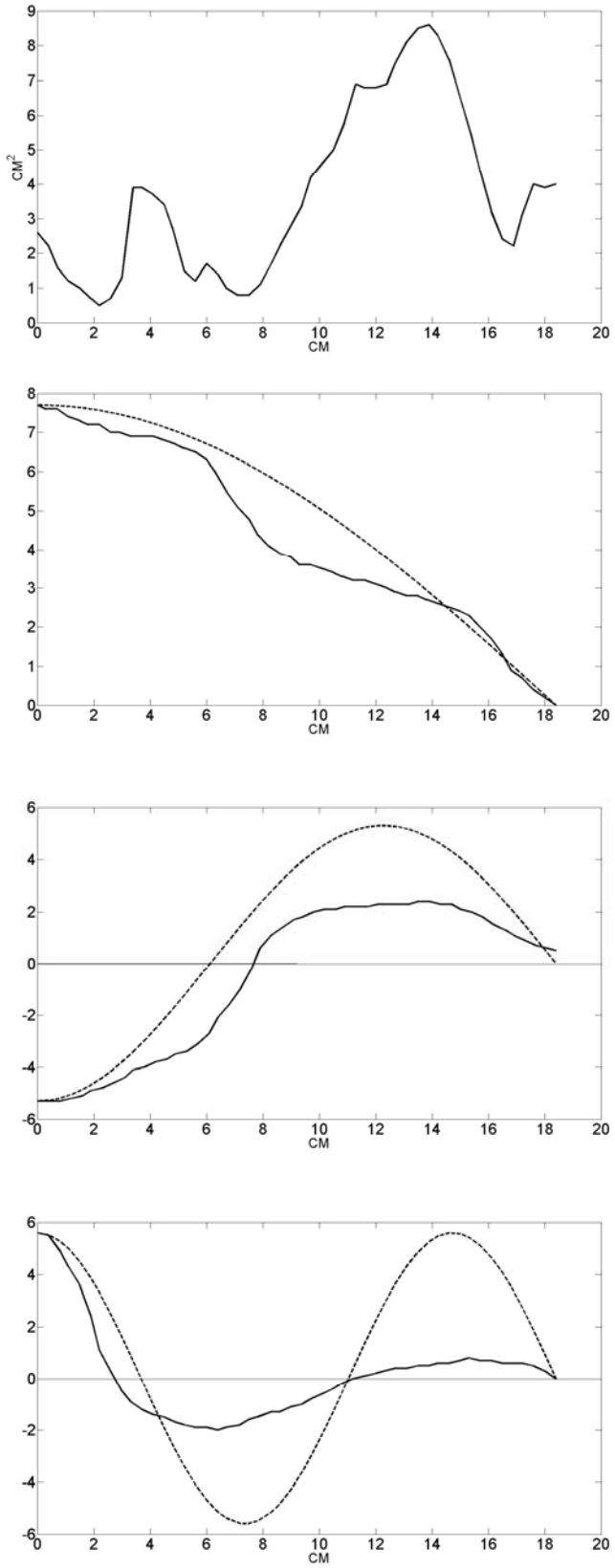


Рис. 2. Площадь поперечного сечения тракта при артикуляции звука /*a*/, собственные функции тракта (─) и собственные функции однородной трубы (---).

$$\eta_i = \frac{\varepsilon\left\{\left[\psi_i^{(0)'}(x_0)\right]^2 - \left[\lambda_i^{(0)}\psi_i^{(0)}(x_0)\right]^2\right\}}{\lambda_i^{(0)^2}},$$

где $\psi_i^{(0)}$ и $\lambda_i^{(0)}$ - собственные функции и их собственные числа для исходной площади речевого тракта. Отсюда видно, что наибольшее влияние на частоту *i*-го резонанса оказывает возмущение, приложенное либо в узле *i*-й собственной функции ($\psi_i^{(0)}(x_0)=0$), либо в ее пучности ($\psi_i'^{(0)}(x_0)=0$). Увеличение или уменьшение частоты *i*-го резонанса возмущенной системы зависит от знака ε. Например, при сужении ($\varepsilon<0$) в первом случае частота резонанса понижается, а во втором - повышается. Частота *i*-го резонанса останется неизменной, когда

$$\left[\psi_i^{(0)}(x_0)\right]^2 = \left[\lambda_i^{(0)}\psi_i^{(0)}(x_0)\right]^2.$$

В том случае, когда возмущение площади тракта приходится как на узел, так и на пучность некоторой собственной функции, то ее собственное число не изменяется. Поэтому подобное возмущение площади мало влияет на резонансы тракта с высокой частотой, и эти частоты зависят только длины тракта так, как если бы это была однородная акустическая труба. На этом свойстве может быть основана оценка длины речевого тракта, если в речевом сигнале измерены резонансные частоты высокого порядка.

На Рис. 2 показаны площадь поперечного тракта, соответствующая его форме на Рис. 1, а также первые три собственные функции тракта и собственные функции однородной трубы такой же длины.

Оценка длины речевого тракта как длины однородной трубы по высокочастотным резонансам связана с определенными трудностями. С одной стороны, чем выше частота резонанса, тем меньше она зависит от формы речевого тракта. С другой стороны, на высоких частотах возникают поперечные волны со своими резонансами, которые трудно отличить от резонансов плоских волн. По оценке [16], плоские волны, для которых справедлива связь между частотой резонанса и длиной трубы (1), распространяются в однородной трубе с жесткими стенками лишь до частот 8 кГц, тогда как, по оценке [6], поперечные волны в речевом тракте возникают уже начиная с частот 4 кГц. К тому же, амплитуда высокочастотных компонент речевого сигнала мала, поэтому подвержена влиянию помех, и зависит от характеристик источника возбуждения. Для некоторых голосов, особенно женских, спектр речевых сигналов настолько ограничен, что на нем отсутствуют резонансы выше третьего.

Еще одна проблема заключается в том, что в спектре речевого сигнала трудно оценить порядковый номер резонанса. В ряде случаев может оказаться предпочтительным не использовать выражение (1), а оценивать длину тракта по разности ближайших формантных частот в высокочастотной области как

$$L = c_0 / 2(F_k - F_{k-1}) \qquad (2)$$

3. ОЦЕНКИ ПО МОДЕЛИ АРТИКУЛЯЦИИ

Для сопоставления различных методов оценки длины речевого тракта по его резонансным частотам необходимо располагать одновременными измерениями этой длины и резонансных частот. В принципе, такие измерения можно было бы выполнить с помощью магнитно-резонансной томографии. Однако такие данные труднодоступны и обладают погрешностью, связанной с толщиной пластин (около 0.5 см). Погрешность оценки резонансных частот по вычисленным формантным частотам в этом методе связана с необходимостью длительной фонации одного и того же гласного, в процессе которой происходят неконтролируемые изменения формы речевого тракта. Ограниченные, но достаточно точные данные можно получить, используя артикуляторный синтезатор речи. В [15] приводятся данные о резонансных частотах вплоть до 6-го резонанса, длине и форме речевого тракта для гласных русского языка. Эти данные приведены в Табл.1 вместе с оценками длины речевого тракта разными методами. Здесь $\Delta F_{56} = F_6 - F_5$ есть

разница между частотами 6-го и 5-го резонансов, а $\bar{F}_2 = (F_1 + F_2 + F_3)/3$ - средняя частота второй форманты.

Таблица 1. Оценка длины речевого тракта по его резонансным частотам.

Гласная	Длина тракта	$L_{\Delta F} = c_0 / 2\Delta F_{56}$		$L_F = 3c_0 / 4\bar{F}_2$		$L=(L_1+L_2+L_3)/3$		L_6	
	см	см	ε, %	см	ε	см	ε, %	см	ε, %
А	19.36	18.64	-3.7	21.44	10.7	19.94	3.0	20.32	5.0
О	19.36	25,09	29.6	21.12	9	21.73	12.2	20.35	5.1
У	21.12	21.28	0.8	23.8	12.7	24.47	15.8	21.70	2.7
И	16.28	16.09	-1.2	13.67	-16	15.51	-4.7	15.31	-6.0
Ы	20.68	25.38	22.7	21.07	19	20.13	-2.6	21.76	5.2
Э	19.80	26.89	35.8	19.3	-2.5	18.93	-4.3	22.77	15.0
Средняя абсолютная ошибка			**15.6**		**8.8**		**7.1**		**6.5**

Оценка длины речевого тракта как средней длины, вычисленной для каждой из трех формантных частот, примерно на 1.7% лучше, чем оценка, полученная для средней частоты второй форманты. Однако в условиях определения формантных частот по речевому сигналу возникают значительные отклонения от реальных частот. Если эти отклонения случайны и независимы, то выражение $L = 3c_0 / 4\bar{F}_2$ может оказаться предпочтительнее, чем $L=(L_1+L_2+L_3)/3$.

Наименьшая ошибка получена при использовании частоты 6-го резонанса, но в реальном речевом сигнале часто трудно определить порядковый номер форманты, а иногда она просто отсутствует. Проблема индексации форманты частично может быть решена использованием разности ближайших формантных частот в области выше 4000 Гц, хотя встречаются и такие голоса, у которых частотный диапазон спектра не достигает этой величины.

Следует отметить, что в этой Таблице представлена только длина тракта как расстояние вдоль средней линии. Концевая поправка, например, на 3%, уменьшает погрешность оценки длины тракта по F_6 для /*А*/ до 2%, /*О*/ - до 2%, /*У*/ - до 0.2%, /*Ы*/ - до 2.1%, /*Э*/ - до 11.6%, и лишь для /*И*/ - ошибка возрастает до -8.7%. При этом средняя абсолютная ошибка падает до 4.4%.

Считается, что, поскольку площадь поперечного сечения тракта при артикуляции гласного /*Э*/ наиболее равномерна, то и оценка длины тракта по резонансным частотам этого звука должна ближе всего соответствовать истинной длине тракта. Однако и у этого звука наблюдается существенное сужение тракта ниже надгортанника, так что ошибка оценки длины тракта по L_6 и $L_{\Delta F}$ в Таблице 1 оказывается максимальной по сравнению с оценками по средним частотам резонансов.

4. ИЗМЕРЕНИЕ ФОРМАНТНЫХ ЧАСТОТ

Определение формантных частот как локальных пиков в амплитудном спектре речевого сигнала до сих пор составляет значительные трудности. Эти трудности связаны как с особенностями генерации звука в речевом тракте, так и с внешними условиями. В результате взаимодействия с акустическими колебаниями в подсвязочной области, в спектре гласных звуков могут либо появляться локальные экстремумы, связанные с резонансами подсвязочной области, либо подавляться резонансные колебания ротовой полости, близкие к резонансам подсвязочной области. Аналогично, неполное перекрытие небной занавеской прохода в носовую полость может привести к появлению антирезонансов и дополнительных пиков в спектре сигнала. Прием речевого сигнала в помещении с реверберацией сопровождается подавлением определенных частот и появлением ложных пиков в спектре. Трудности в оценке резонансных частот усугубляются при использовании телефонов.

Существует множество алгоритмов оценки резонансных частот речевого тракта. Среди них наиболее часто используются частоты пиков амплитудного спектра, полюса передаточной

функции тракта, найденные по методу линейного предсказания, и интервалы между пересечениями нулевого значения сигнала в определенных частотных полосах [17]. Все методы обладают преимуществами и недостатками в зависимости от условий приема речевого сигнала, типа голоса диктора и конкретного фонетического состава речи.

В настоящей работе поиск формантных частот осуществлялся с использованием гребенки фильтров, свойства которых примерно соответствуют свойствам базилярной мембраны в слуховом анализаторе человека. Эти фильтры были предложены в [12], и получили название gammatone фильтров. Во временной области отклик каждого такого фильтра описывается как

$$g(t) = t^{n-1}e^{-bt}\cos(\omega t + \varphi)$$

где n – порядок функции (обычно n=4), b – определяет ширину полосы пропускания, ω - центральная круговая частота , а φ - фазовая константа, которая обычно принимается равной нулю. Преобразование Лапласа для такого фильтра есть [9]

$$G(s) = \frac{6(-b^4 - 4b^3 s - 6b^2 s^2 - 4bs^3 - s^4 + 6b^2\omega^2 + 12bs\omega^2 + 6s^2\omega^2 - \omega^4)}{(b^2 + 2bs + s^2 + \omega^4)^4}$$

В системе gammatone шкала частот может быть выбрана произвольно, что создает гибкость в разработке методов анализа речи.

На Рис. 3 показан аналог динамической спектрограммы, полученной путем вычисления огибающей положительных значений сигнала для каждого фильтра. Использование только положительных значений сигналов обосновано тем фактором, что внутренние волосковые клетки Кортиева органа отвечают только на положительные смещения базилярной мембраны.

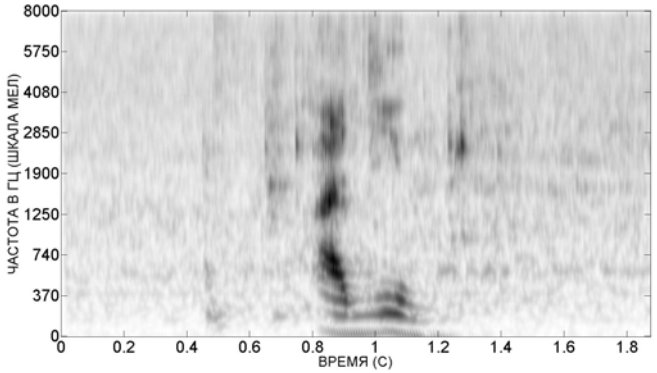


Рис. 3. Спектрограмма слова "один".

На Рис. 4 показан спектральный разрез в момент времени t = 0.845

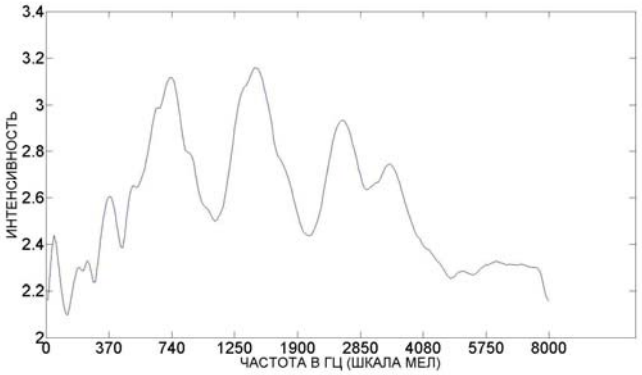


Рис. 4. Амплитудные значения спектрограммы.

Система gammatone фильтров обеспечивает высокое разрешение по частоте, что является преимуществом при оценке частот локальных максимумов спектра на относительно высоких частотах. Однако одновременно с этим на низких частотах происходит разделение гармоник основного тона, что маскирует положение первой форманты, поскольку пик энергии всегда приходится на какую-либо гармонику. Для того, чтобы избежать этого явления при оценке частоты первой форманты, было выполнено кепстральное преобразование. Коэффициенты дискретного кепстрального преобразования вычисляются как

$$c_n = \sum_{m=1}^{M} [\log Y(m)] \cos[\frac{\pi n}{M}(m - \frac{1}{2})],$$

где $Y(m)$ – выходной сигнал m-го фильтра, c_n – n-й коэффициент кепстра, M - число фильтров.

После того, как число коэффициентов было уменьшено до 7 (для оценки частоты первой форманты) или 11 (для оценки частоты низкой второй форманты), выполнялось обратное кепстральное преобразование, и на полученных спектрограммах отыскивались локальные пики энергии.

В [10] было показано, что задача определения резонансных частот речевого тракта является некорректной, и для ее решения необходима дополнительная информация. Поскольку диапазон значений первых трех формант для гласных русского языка известен, то можно подсчитать долю неправильно найденных формант при непосредственной оценке формантных частот по спектрограмме. Такая доля составляет около 30%. Если использовать совместное вероятностное распределение трех формант в русском языке, то ошибка остается довольно высокой - около 20%. Свести ошибку оценки форматных частот до величин порядка 5% удалось только в том случае, когда использовалось распределение формант для каждого гласного порознь.

На Рис. 5 синхронно с отсчетами частоты основного тона показаны дискретные оценки частот формант гласного /*a*/ в слове "*два*". Частота формант находилась как среднее значение наиболее часто представленных дискретных оценок на заданном интервале времени.

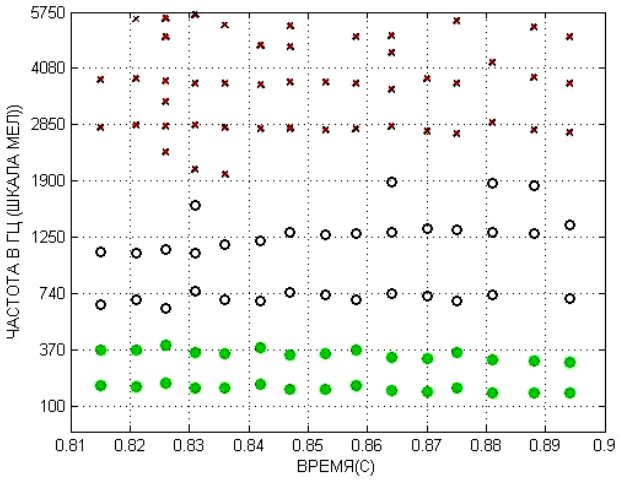


Рис. 5. Треки формантных частот ударного гласного /*a*/ в слове "*два*". (xxx) оценки по исходной спектрограмме, (ooo) - оценки после кепстральных преобразований, (•••) основной тон и его первая гармоника.

5. РЕЗУЛЬТАТЫ ЭКСПЕРИМЕНТОВ

1. Распознавание пола.

Эксперименты проводились с использованием базы данных для однозначных числительных русского языка. Речь 216 мужчин и 177 женщин записывалась в различных

акустических условиях через различные микрофоны и разные преобразователи "аналог/цифра". Всего было представлено 285361 произнесений. В силу неустойчивости определения частоты 6-й форманты оценка длины тракта выполнялась либо по разным частотам высших формант, либо по первым трем формантам.

Первый этап состоял в оценке длины речевого тракта для мужского и женского пола.

На рис. 6 показано распределение длин тракта для мужчин и женщин, вычисленное по разности высших формант для всех произнесений числительных.

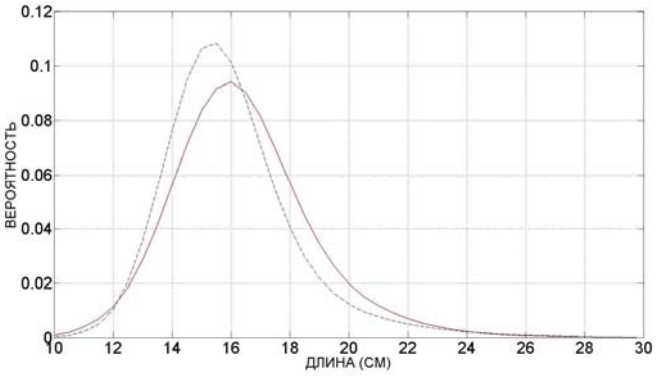


Рис. 6. Распределение длин речевого тракта мужчин (—) и женщин (---) по разности высших частот.

Как видно, наиболее вероятная длина тракта у мужчин (16 см), в среднем, несколько больше, чем у женщин (15.2 см), и все распределение длин у мужских голосов сдвинуто в область больших длин тракта. Это вполне соответствует анатомическим данным, хотя эти оценки кажутся заниженными. Средняя длина тракта у мужчин (16.39 см) и средняя длина у женщин (15.97 см) представляются вполне правдоподобными, но минимальная длина в 10 см и максимальная длина в 29 см, возможно, выходят за реальный диапазон анатомических размеров взрослых людей.

Ошибка распознавания пола по этому параметру для всех произнесений (т.е. независимо от контекста) составляет около 87%. Минимальная ошибка распознавания пола достигается на ударных гласных слов "*пять*" и "*девять*" (83% и 84%), тогда как ошибка, усредненная по оценкам каждого слова составляет примерно 86.5%, а максимальная ошибка (90.2%) найдена для ударного гласного /*а*/ в слове "*два*".

На Рис. 7 показано распределение оценок длины тракта мужчин и женщин, полученное по всем произнесениям при использовании частот первых трех формант.

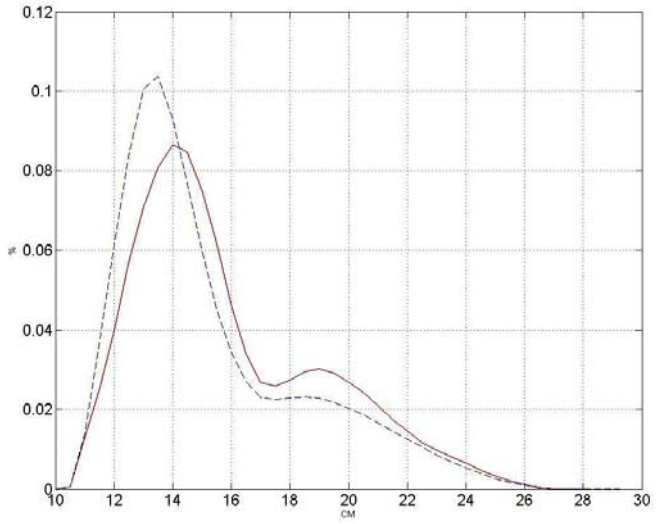


Рис. 7. Распределение длин речевого тракта у мужчин (—) и женщин (---) по средней частоте второй форманты.

Здесь, как и на Рис. 6, видно, что наиболее вероятная длина тракта у мужчин больше, чем у женщин, но наиболее вероятные длины тракта уменьшились примерно на 2 см - до 14 см у мужчин, и 13.5 см у женщин. Средние значения длины тракта составляют 15.85 см у мужчин, и 15.27 см - у женщин, а диапазон оценок несколько сократился - от 10.5 см до 28.5 см. Смещение наиболее вероятной длины тракта в сторону меньших значений объясняется тем, что в словаре однозначных числительных доминируют слова с ударными гласными /*и*, *ы*, *я*, *е*/, у которых средняя частота второй форманты $\overline{F}_2$ высока. Второй пик оценок в области 19 см соответствует словам "*два*, *шесть*", ударные гласные которых имеет относительно низкую вторую форманту. Это указывает на зависимость оценки длины речевого тракта от конкретного гласного звука.

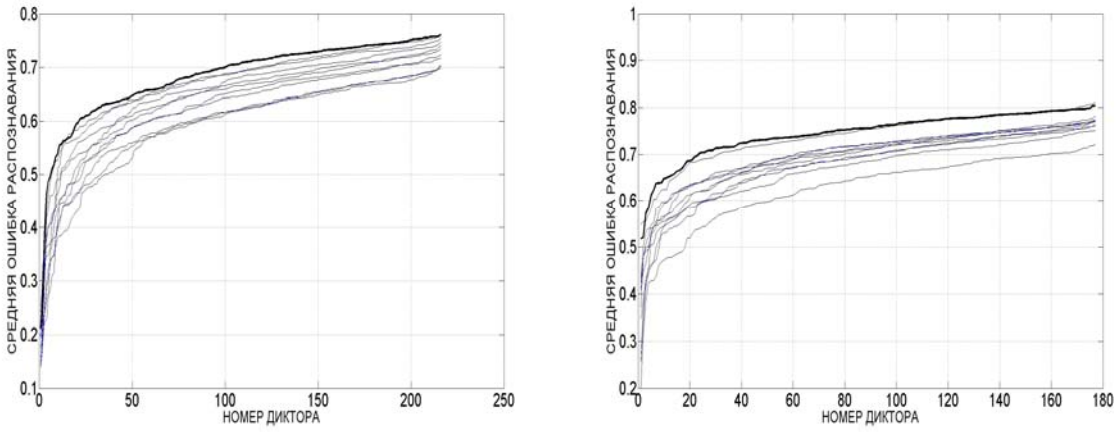
Ошибка распознавания пола по этому параметру независимо от контекста составляет около 87.7%. Минимальная ошибка распознавания пола достигается на ударных гласных слов "*четыре*", "*пять*" и "*девять*" (около 82%), тогда как ошибка, усредненная по оценкам каждого слова, составляет примерно 86.5%, а максимальная ошибка (94.5%) найдена для ударного гласного /*о*/ в слове "*восемь*".

Из выполненных экспериментов следует, что ошибка распознавании пола при оценках длины тракта и по разности высших формант, и по первым трем формантам, почти одинакова, составляя примерно 87% ÷ 88%.

2. Распознавание диктора.

Распознавание диктора по длине речевого тракта, в отличие от распознавания пола, демонстрирует существенно другие характеристики.

На Рис. 8а, б показаны средние ошибки распознавание дикторов в зависимости и независимо от контекста и ошибки распознавания для конкретных гласных при оценке длины тракта по разности высших формантных частот ΔF_{56}.



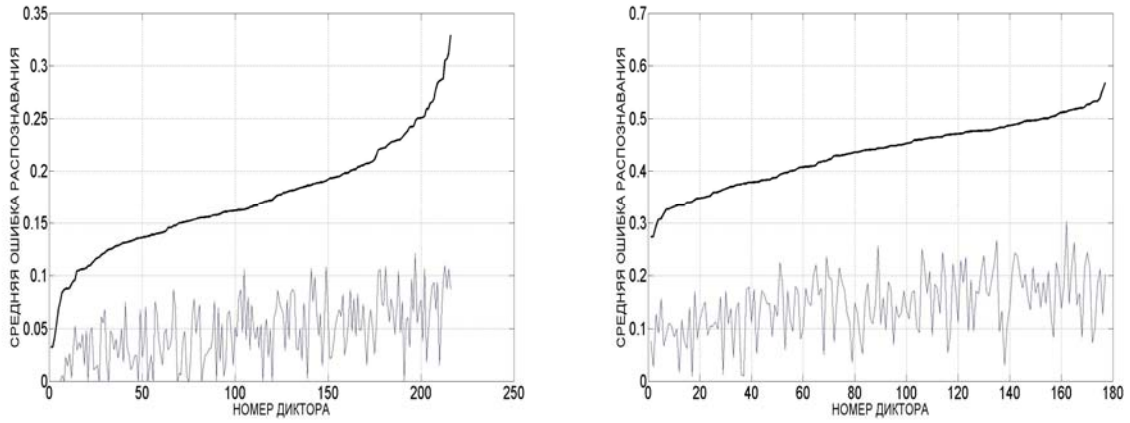
а) б)

Рис. 8. Ошибка распознавания мужчин (а) и женщин (б). Верхняя кривая– средняя ошибка независимо от контекста, нижние кривые – для каждого гласного. По оси абсцисс - номер диктора.

Видно, что и у мужчин, и у женщин имеется небольшое количество дикторов, относительно хорошо различающихся по этому параметру. При этом средняя ошибка распознавания дикторов при оценке длины тракта независимо от контекста составляет для мужчин около 68.3%, а для женщин - примерно 74.5%.Усредненная ошибка по отдельным гласным оказывается несколько ниже - 63.1% для мужчин, и 68.8% - для женщин. Иными словами, выигрыш при контекстно-зависимом анализе составляет, в среднем, около 5%. Наименьшие ошибки распознавания для мужчин достигаются на ударных гласных слов "*пять*, *девять*", где средняя ошибка составляет 59.4 % и 59.1%. Аналогично, наименьшие ошибки распознавания для женщин достигаются на ударных гласных слов "*два*, *пять*", где средняя ошибка составляет соответственно 62.6 % и 66.4%.

Влияние контекстно-зависимого анализа проявляется при поиске пар дикторов, отличающихся лучше всего. На Рис. 9а, б показаны минимальные ошибки распознавания для

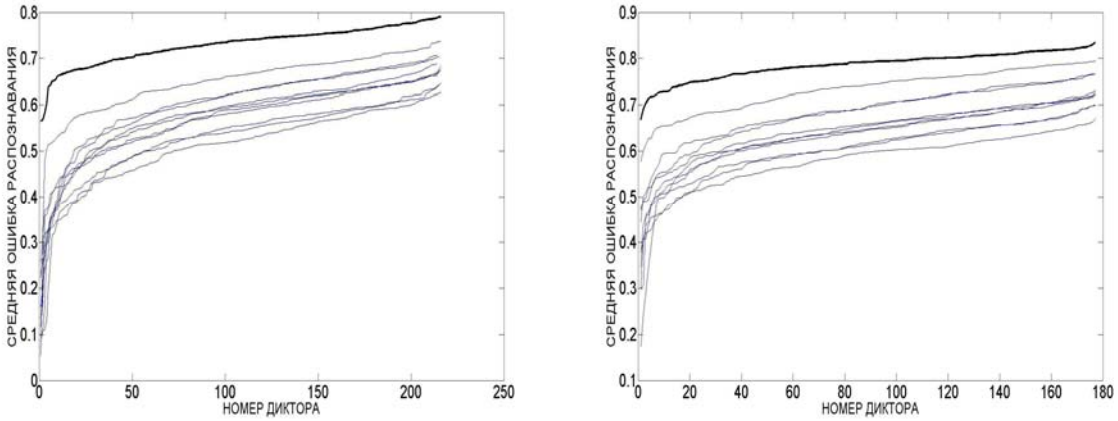
каждого диктора, т.е. такие ошибки, которые достигаются хотя бы для одной пары сопоставляемых дикторов. Нумерация дикторов установлена такими образом, чтобы ошибка при контекстно-независимом распознавании монотонно возрастала (верхняя кривая). Минимальные ошибки распознавания представлены на нижней кривой для той же последовательности номеров дикторов, что и для верхней кривой. Видно, что при контекстно-независимом распознавании для каждого диктора всегда найдется хотя бы один конкурент, похожий на него с вероятностью ниже примерно 15% для мужчин и 25% - для женщин.



а) б)

Рис. 9. Минимальная ошибка распознавания мужчин (а) и женщин (б). Верхняя кривая – средняя ошибка по всем произнесениям, нижние кривые – для каждого гласного. По оси абсцисс - номер диктора.

При распознавании дикторов с оценкой длины тракта по первым трем формантным частотам средняя ошибка при контекстно-независимом анализе для мужчин составила 73%, а для женщин - 78.5%, что значительно хуже, чем при анализе по разности высших формант. В то же время, контекстно-зависимый анализ демонстрирует существенное снижение ошибок распознавания: в среднем достигается 57.1% для мужчин, и 64.5% - для женщин (Рис. 10). При этом минимальные ошибки распознавания при контекстно-зависимом анализе, как для мужчин, так и для женщин для каждого диктора оказались близки к нулю.



а) *б*)

Рис. 10. Ошибка распознавания мужчин (а) и женщин (б) для оценок по средней форманте. Верхняя кривая – средняя ошибка по всем произнесениям, нижние кривые – для каждого гласного.

Средняя разность наиболее вероятной длины тракта у мужчин по всем произнесениям составляет примерно 1.86 см, а у женщин - 1.98 см. Наименьшие ошибки распознавания для

мужчин достигаются на ударных гласных слов "*ноль*, *четыре*", где средняя ошибка составляет 52.7% и 51.3%. Аналогично, наименьшие ошибки распознавания для женщин достигаются на ударных гласных слов "*семь*, *девять*", где средняя ошибка составляет соответственно 58.1% и 60.2%.

В Табл. 2, 3 сопоставляются оценки длины тракта по разности высших формант и первым трем формантам для ударных гласных разных слов для отдельного мужчины и женщины.

Таблица 2. Наиболее вероятная длина тракта. Диктор мужчина.

Слово	$L_{\Delta F} = c_0 / 2\Delta F_{56}$ см	$L_F = 3c_0 / 4\overline{F}_2$ см	Разность оценок $L_F - L$ см
ноль	16.5	19.5	3
один	16	14.5	-1.5
два	15.5	18	2.5
три	15.5	14	-1.5
четыре	16	12.5	-3.5
пять	16.5	15.5	-1
шесть	16.5	15.5	-1
семь	17	14.5	-2.5
восемь	16	20.5	4.5
девять	16.5	13	-3.5
Среднее	**16.2**	**14.4**	**2.2**

Из этой Таблицы следует, что наименьшее разногласие в оценках наиболее вероятной длины тракта при контекстно-зависимом анализе для гласных /*я*, э/ в словах "*пять*, *шесть*".

Таблица 3. Наиболее вероятная длина тракта. Диктор женщина.

Слово	$L_{\Delta F} = c_0 / 2\Delta F_{56}$ см	$L_F = 3c_0 / 4\overline{F}_2$ см	Разность оценок $L_{\Delta F} - L_F$ см
ноль	15.5	19	3.5
один	15	14	-1
два	15.5	16.5	-1
три	15.5	13	-2.5
четыре	15.5	13.5	-2
пять	15.5	14	-1.5
шесть	15.5	14	-1.5
семь	15.5	12.5	-3
восемь	15.5	18.5	3
девять	15.5	12	-3.5
Среднее	**15.4**	**14.7**	**2.2**

Здесь наименьшее разногласие в оценках наиболее вероятной длины тракта при контекстно-зависимом анализе для гласных /*и*, *а*/ в словах "*один*, *два*".

6. ОБСУЖДЕНИЕ И ЗАКЛЮЧЕНИЕ

Оценка длины речевого тракта как по усредненным частотам первых трех формант, так и по разности высших формант должна сопровождаться значительными вариациями в силу разной высоты гортани, огубления и формы тракта для разных гласных. Это наглядно проявилось в экспериментах при контекстно-независимом подходе, где вероятность правильного распознавания

диктора находится в диапазоне 25% - 35% при оценке по высшим формантам, и еще ниже (около 10%) при анализе по первым трем формантам.

В базе данных, использованной в описанных выше экспериментах, доминируют мягкие гласные, для которых характерен подъем гортани при артикуляции. Поэтому можно ожидать смещение длины тракта в область меньших значений, а также более плохую различимость как пола диктора, так и голосов внутри каждого пола. Более фонетически сбалансированная база данных вряд ли существенно улучшит вероятность правильного распознавания пола, которая в наших экспериментах составляет всего около 13%. Даже при контекстно-зависимом распознавании, когда известен тип гласного, вероятность распознавания пола повышается всего до 17%. Это весьма малая величина по сравнению с известными по литературе показателями. Так, в [17] вероятность правильного распознавания пола по одному гласному составляет около 90%. Вместе с тем, этот слабый признак может оказаться полезен в совокупности с другими параметрами.

Оценка длины речевого тракта представляется более перспективной в задачах распознавания диктора, особенно в условиях независимости от контекста, где вероятность правильного распознавания мужских голосов с оценкой длины тракта по разности частот высших формант составила около 31 %, а женских голосов - около 25%. В случае отсутствия высших резонансов в спектра речевого сигнала надежнее всего вообще отказаться от этого параметра. Если же и нижние формантные частоты определены с ошибкой, которая приводит к значительному отклонению от диапазона оценок длины тракта для данного диктора, полученных на этапе обучения, то также следует отказаться от использования этого параметра.

При контекстно-зависимом распознавании, когда заранее известен тип гласного, вероятность правильного распознавания мужских голосов с оценкой длины тракта по первым трем формантам повышается до 43 %, а женских - до 35.5%. При таких условиях длина тракта становится достаточно сильным признаком для использования совместно с другими признаками, такими, как, например, параметры модели голосового источника.

Литература

1. Atal B. Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identfication and verification. *J.Acoust.Soc.America*, 1974, vol. 55, 1304–1312.
2. Bachorowski J., Owren M.J. Acoustic correlates of talker sex and individual talker identity are present in a short vowel segment produced in a running speech. Journ. Acoust. Soc. Am., 1999,vol. 106., 1054-1063.
3. Dusan S. Estimation of speaker's height and vocal tract length from speech signal. Proc. of EUROSPEECH'2005. Lisbon, Portugal, 1989-1992.
4. Dusan S., Deng Li. Vocal tract length normalization for acoustic-to-articulatory mapping using neural nets. Journ. Acoust. Soc. Am., 1999, vol. 106, N4, Pt. 2, 2181.
5. Eide E., Gish H. A parametric approach to vocal tract length normalization. In Proceedings of the International Conference on Acoustics, Speech and Signal Processing, 1996, 346-348.
6. Фланаган Д. Анализ, синтез и восприятие речи. 1968. Связь, 392 с.
7. Fitch W. T., Giedd J. Morphology and development of the human vocal tract: A study using magnetic resonance imaging. Journ. Acoust. Soc. Am., 1999, vol. 106, N3, Pt. 1, 1511 – 1522.
8. Lee L., Rose R. A frequency warping approach to speaker normalization. IEEE Trans. Acoust., Speech, Signal Proc., 1998, vol. 6, N1, 49–60.
9. Leng Y. R., Tran H. D., Kitaoka N., Li H. Selective gammatone filterbank feature for robust sound event recognition. Interspeech 2010, 2246-2249.
10. Леонов А.С., Сорокин В.Н. Об однозначности определения резонансных частот голосового тракта по речевому сигналу. Доклады Академии Наук, 2011, т. 440, №1, 32-34.
11. Naito M., Deng Li, Sagisaka Y. Speaker clustering for speech recognition using vocal tract parameters. Speech Communication, 2002, vol. 36, 305-315.
12. Patterson R. D., Holdsworth J. A functional model of neural activity patterns and auditory images. Advances in Speech, Hearing and Language Processing, 1996, vol. 3, 547–563.
13. Pitz M., Ney H. Vocal tract normalization equals linear transformation in cepstral space. *IEEE Trans. Speech and Audio Processing*, 2005, vol. 13, N5, 930–944.
14. Сорокин В.Н. Теория речеобразования. Радио и связь, М. 1985. 313 с.

15. Сорокин В.Н. Синтез речи. Наука, М. 1992. 392 с.
16. Сорокин В.Н. Речевые процессы. Народное образование, М. 2012. 599 с.
17. Сорокин В.Н., Макаров И.С. Распознавание пола диктора по голосу. Акустический ж., 2008, т. 54, №4, 1- 9.
18. Turner RE, Al-Hames MA, Smith DRR, Kawahara H, Irino T, Patterson RD. Vowel normalisation: Time-domain processing of the internal dynamics of speech. In: Divenyi P, editor. Dynamics of Speech Production and Perception. IOS Press, 2005. 153-170.
19. Zhan P., Westphal M. Speaker Normalization Based on Frequency Warping. In Proceedings of the International Conference on Acoustics, Speech and Signal Processing, 1997. 1039-1042.

Estimation of vocal tract length.
V.N.Sorokin, I.V.Geraskin

Abstract

We studied two methods of estimating the length of the vocal tract, the equivalent length of a uniform acoustic tube. One method was based on calculating the tract length by difference between the closest frequencies of the local maxima of the spectrum, which are higher than 4 kHz. In another method, tract length was calculated from the average frequency of the second formant frequencies determined by the first three formants. Also considered options regardless of the context analysis, and depending upon known vowels.The probability of correct recognition of gender speaker by both methods was about 13%, and is almost independent of knowledge of the context. The probability of correct recognition of men's voices from the difference between the higher formants independently of the context was about 31%, female voices - about 25.5%, and depending on the context - 37% and 31%, respectively. The context-independent recognition by the average frequency of the second formant of male voices is close to 27%, and female voices - to 21.5%, whereas in the context-dependent mode of a known type of vowel probability of correct recognition of men's voices raised to 43%, and female voices - up to 35.5%.

Key words: the length of the vocal tract, gender recognition, speaker recognition.