

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

НАЧАЛЬНЫЕ УСЛОВИЯ В ЗАДАЧЕ ИДЕНТИФИКАЦИИ ГОЛОСОВОГО ИСТОЧНИКА

В.Н.Сорокин, А.А.Тананыкин

Институт проблем передачи информации, Российская академия наук, Москва, Россия

Поступила в редколлегию 14.01.2010

Аннотация. В задаче идентификации говорящего по параметрам голосового источника выполняется аппроксимация вычисленного методом линейного предсказания сигнала-остатка математической моделью источника путем минимизации среднеквадратической ошибки. Результаты такой аппроксимации зависят от начальных условий, поиск которых представляет собой самостоятельную задачу. Эта задача решалась путем кластеризации на множестве данных, полученных для начальных условий произвольного вида. Исследовались различные меры сходства и различные способы поиска центров кластеров. Наилучшей метрикой оказалась Евклидова. Центры кластеров для мужчин и женщин заметно различаются. Не обнаружено явной зависимости центров кластеров от частоты основного тона.

Голосовой источник возбуждения акустических колебаний в речевом тракте пропорционален производной от объемной скорости воздушного потока через голосовую щель. Параметры этого источника можно использовать для решения таких задач, как распознавание речи, в частности, определение пола диктора [1], создание информативных признаков при верификации диктора [2], в задаче синтеза речи с имитацией разной индивидуальности, а также при диагностике патологии гортани.

Основным методом поиска голосового источника является обратная фильтрация с использованием линейного предсказания. Этот метод основан на предположении линейности системы “голосовой источник - речевой тракт”. В такой модели речевой тракт не оказывает влияния на голосовой источник, и их можно рассматривать в отдельности независимо друг от друга, а спектр речевого сигнала представляет собой произведение передаточной функции речевого тракта на спектр голосового источника. Построив модель речевого тракта, методом обратной фильтрации можно попытаться вычислить голосовой источник по речевому сигналу. Обычно предполагается, что передаточная функция речевого тракта имеет только полюса и записывается в виде:

$$A(z) = \frac{1}{1 + \sum_{k=1}^p a_k z^{-k}}$$

где p – порядок модели [3]. После того, как по речевому сигналу найдены коэффициенты a_k , производится процедура обратной фильтрации, состоящая в том, что речевой сигнал пропускается через фильтр с передаточной характеристикой $1/A(z)$. Нули этого фильтра совпадают с полюсами модели речевого тракта и, таким образом, при фильтровании сигнала происходит его восстановление до состояния перед прохождением сигнала через речевой тракт. Полученный сигнал-остаток R рассматривается как сигнал голосового возбуждения акустических колебаний в речевом тракте.

Вследствие погрешностей определения частот и затуханий полюсов передаточной функции $A(z)$, процедура обратной фильтрации не гарантирует точного вычисления сигнала возбуждения. При прослушивании этого сигнала можно уловить содержание речевого высказывания, что свидетельствует о неполном режектировании характеристик речевого тракта. К тому же, метод линейного предсказания чувствителен к помехам и искажениям в канале связи. Поэтому во многих случаях сигнал-остаток подвержен искажениям, которые необходимо компенсировать при определении параметров голосового источника. Примеры удачного вычисления сигнала остатка и искаженного сигнала показаны на Рис. 1 и 2.

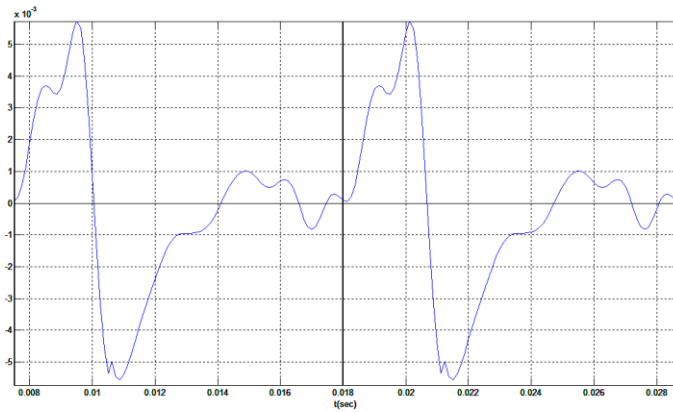


Рис. 1. Сигнал-остаток на двух периодах основного тона. Сплошными вертикальными линиями обозначены условные границы периода основного тона.

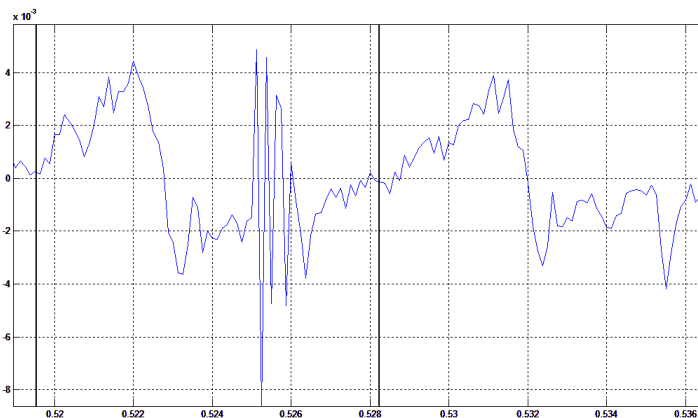


Рис. 2. Сигнал-остаток на двух периодах основного тона. Сплошными вертикальными линиями обозначены условные границы периода основного тона.

Особенности обратной фильтрации приводят к тому, что задача определения параметров голосового источника является некорректной, т.е. не гарантирует однозначного и устойчивого решения. Для ее решения необходимо использовать дополнительную информацию о возможной форме импульса и ограничениях на его параметры.

В качестве приблизительной формы импульса хорошо зарекомендовала себя параметрическая модель, описанная в [4]:

$$w'_m(t) = \begin{cases} A_1 \sin\left(\frac{\pi t}{2T_1}\right), & 0 \leq t \leq T_1 \\ (A_1 + A_2) \cos\left(\frac{\pi(t-T_1)}{2(T_2-T_1)}\right) - A_2, & T_1 < t \leq T_2 \\ -A_2 \left(\frac{T_3-t}{T_3-T_2}\right)^2, & T_2 < t \leq T_3 \\ 0, & T_3 < t \leq T_0 \end{cases}$$

Форма импульса голосового возбуждения, которая порождается этой моделью, показана на Рис. 3.

Интервал времени $[0, T_1]$ соответствует участку, на котором голосовая щель открывается, и ускорение воздушного потока возрастает от нуля до максимального значения A_1 . На интервале времени $[T_1, T_2]$ скорость потока проходит через свой максимум, а ускорение потока уменьшается до отрицательной величины A_2 . На интервале $[T_2, T_3]$ скорость потока продолжает падать, и в момент времени T_3 его скорость

и ускорение становятся равными нулю. На интервале $[T_3, T_0]$ голосовая щель закрыта, и T_0 есть период основного тона.

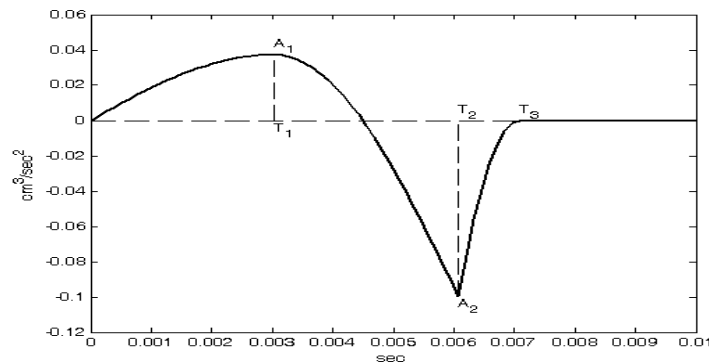


Рис. 3. Производная от объемной скорости источника голосового возбуждения для периода основного тона $T_0 = 10$ мс.

В этой модели шесть параметров $T_1, A_1, T_2, A_2, T_3, T_0$, причем T_0 не участвует в процессе оптимизации, поскольку период основного тона определяется отдельным детектором, описанным в [5]. Конечная цель процесса идентификации голосового источника состоит в вычислении площади голосовой щели по объемной скорости потока [1, 2, 4], и для этого нужно проинтегрировать вычисленное ускорение $w'_m(t)$. При этом в общем случае возникает некая постоянная интегрирования, не имеющая физического смысла. Поэтому приходится наложить условие

$$\int_0^{T_0} w'(t) dt = 0,$$

которое связывает два параметра - A_1 и A_2 ,

$$A_1 = A_2[(6 - 3\pi)T_1 + (2\pi - 6)T_2 + \pi T_3] / 6T_2,$$

так что объёмные скорости воздушного потока в начале импульса и в его конце становятся равными нулю.

Параметр A_2 был выбран в качестве независимого параметра, а A_1 искался из описанного выше условия. Параметры T_1, T_2, T_3 нормировались к периоду основного тона T_0 , принимая значения от нуля до единицы. В итоге, в данном исследовании рассматривались четыре параметра T_1, T_2, T_3, A_2 , которые определялись путем минимизации невязки между сигналом-остатком $R(t)$ и $w'_m(t)$ методом наименьших квадратов на периоде основного тона

$$\min \|R(t) - w'_m(t)\|^2$$

Для того, чтобы повысить вероятность нахождения глобального минимума невязки, необходимо выполнить несколько циклов оптимизации с разными начальными условиями, и выбрать такие параметры модели, которые доставляют наименьшее значение невязки. Оптимизация занимает значительную долю времени вычислений. Правильно выбранные начальные приближения также ускоряют сходимость процесса оптимизации, что важно для экономии времени вычислений в практических задачах речевых технологий. В данной работе оптимизация выполнялась в среде MATLAB с помощью функции `lsqcurvefit`.

В качестве материала для получения начальных приближений были использованы записи говорящих мужского и женского полов. Запись высказываний дикторов обоих полов проводилась в комфортных условиях через 4 разных типа микрофонов. Четыреста человек произносили числительные от нуля до девяти через микрофоны четырёх разных типов. Из этих записей по акустическим параметрам с помощью *ЕМ* (*Expectation Maximization*)-алгоритма были выбраны наиболее типичные представители: 49 мужчин и 37 женщин. Записи этих представителей подверглись ручной сегментации на фонемы опытными фонетистами. Из полученного набора фонем для каждого произношения были выбраны ударные гласные звуки, на которых вычислялись параметры голосового источника.

Начальные приближения определялись в два этапа. На первом этапе проводилась визуальная оценка параметров модели голосового источника по форме сигнала остатка. Таким же образом оценивался диапазон возможных значений этих параметров. Для каждого параметра было выбрано два значения, что определило число начальных условий "нулевого приближения", равное 16 (Табл. 1). Эти начальные приближения составили кодовую книгу для оптимизации, выполненной для всего речевого материала.

Таблица 1. Параметры "нулевых" начальных приближений для мужчин и женщин.

№ приближения	T1	A2	T2	T3
1	0.20	0.02	0.47	0.75
2	0.20	0.02	0.47	0.90
3	0.20	0.02	0.70	0.75
4	0.20	0.02	0.70	0.90
5	0.20	0.05	0.47	0.75
6	0.20	0.05	0.47	0.90
7	0.20	0.05	0.70	0.75
8	0.20	0.05	0.70	0.90
9	0.30	0.02	0.47	0.75
10	0.30	0.02	0.47	0.90
11	0.30	0.02	0.70	0.75
12	0.30	0.02	0.70	0.90
13	0.30	0.05	0.47	0.75
14	0.30	0.05	0.47	0.90
15	0.30	0.05	0.70	0.75
16	0.30	0.05	0.70	0.90
Ожидаемый диапазон	0.1-0.4	0.01-1.0	0.35-0.8	0.5-0.9

Проекция распределения параметров источника возбуждения для исходных начальных приближений показаны на Рис. 5. Видно, что некоторые компоненты векторов "нулевых" начальных приближений находятся вне области существования найденных параметров голосового источника.

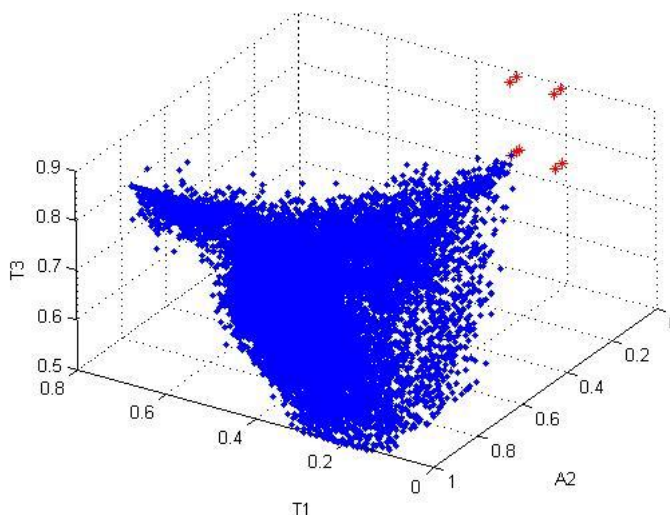


Рис. 5. Проекция распределения параметров источника возбуждения на трёхмерное подпространство. Синими точками выделены вектора базы данных, красными звёздочками - выбранные эвристическим путём 16 векторов начальных приближений.

На втором этапе для каждого пола – мужчин или женщин в отдельности, выполнялась кластеризация вычисленных параметров голосового источника с целью поиска наиболее характерных представителей голосов.

Существует большое число различных способов кластеризации. К ним относятся иерархические алгоритмы, представителями которых является алгоритм К-средних (*K-means*) и его модификации, статистические алгоритмы, основанные на применении *ЕМ*-алгоритма, нейронная сеть Кохонена,

эвристические алгоритмы (ФОРЕЛЬ и другие) и т.д. В нашей постановке задачи нет необходимости описания плотности вероятности на множестве вычисленных параметров голосового источника. К центрам кластеров предъявляется лишь требование достаточно разумного покрытия этого множества. Поэтому в качестве метода кластеризации был выбран классический алгоритм К-средних, реализованный в функции MATLAB *kmeans*. В этом алгоритме минимизируется сумма расстояний между элементами кластера и его центроидом.

Процесс кластеризации проводился для трёх типов мер и двух способов генерации начальной выборки. Объекты в многомерном пространстве рассматриваются как векторы. В качестве меры близости между векторами были выбраны метрики L_2 , L_1 и косинусная мера.

При кластеризации с метрикой L_2 , центроидом является среднее между объектами, образующими кластер:

$$d_{L_2}(x, y) = \sqrt{\sum_{i=1}^n (x_i - y_i)^2}$$

При кластеризации с метрикой L_1 , которая представляет собой сумму абсолютных отклонений, центроидом является медиана наблюдений, образующих кластер:

$$d_{L_1}(x, y) = \sum_{i=1}^n |x_i - y_i|$$

Наконец, косинусная мера определяется как единица минус косинус от угла между объектами, образующими кластер:

$$d_{sc}(x, y) = 1 - \frac{(x, y)}{\|x\| \cdot \|y\|}$$

Алгоритм кластеризации должен быть дополнен способом определения числа кластеров. С этой целью на каждой итерации заданное количество кластеров увеличивалось до тех пор, пока в результате кластеризации не появится кластер, размер которого меньше некоторого заданного порога. В этом случае устанавливается число кластеров, найденное на предыдущем шаге. В качестве такого порога была выбрана величина, равная трём процентам от количества всех векторов в базе говорящих данного пола.

Опыт предыдущих работ показал, что некоторые параметры центроидов кластеров могут выпадать из экспериментально найденного распределения, если это множество не является выпуклым. Поэтому в качестве векторов начального приближения выбираются не центроиды, а ближайшие к ним реальные векторы, которые далее в тексте сохраняют название "центроиды".

Другой проблемой метода кластеризации *K-means* является выбор начальных координат центроидов кластеров. Эта проблема тем более существенна, что сам алгоритм допускает попадание решения в локальный минимум. Поэтому, для каждой выбранной меры на базе данных проводилось по два эксперимента. В первом эксперименте начальные координаты центроидов кластеров выбирались случайным образом из всей базы. Это соответствует способу 'sample' в функции MATLAB *kmeans*. Во втором эксперименте начальные координаты центроидов также выбирались случайным образом, но уже из множества векторов, которое составляет 10% от размера всей базы. Это реализуется при выборе способа определения начальных условий 'cluster' функции MATLAB *kmeans*. Сравнение результатов кластеризации по этим выборкам позволяет оценить степень устойчивости метода относительно начальной выборки.

На этом этапе выяснилось, что формы импульсов голосового источника и объемной скорости для метрики L_2 и косинусной меры устойчивы относительно способа задания исходной выборки, тогда как результат кластеризации в метрике L_1 существенно зависит от исходной выборки. Из Рис. 6 видно, что в косинусной мере практически для всех центроидов интервал открытой голосовой щели равен 0.6 от периода основного тона. И хотя это соответствует общим представлениям о соотношении интервалов открытой и закрытой голосовой щели, такое единообразие вызывает сомнение в том, что найденные центроиды достаточно представительны для разнообразия мужских голосов.

Аналогичные выводы вытекают и из анализа форм импульсов для женских голосов. Отсюда следует, что, в рамках проведенного исследования, только метрика L_2 подходит для поиска начальных приближений методом кластеризации.

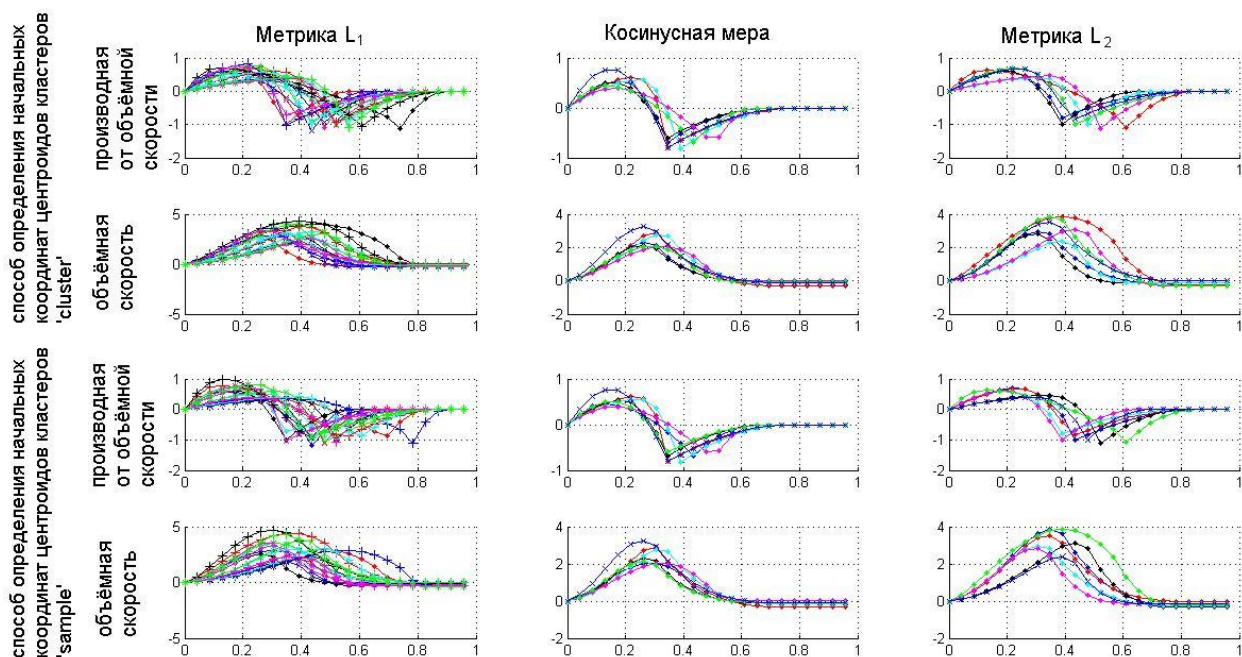


Рис. 6. Импульсы голосового источника и объемной скорости, соответствующие центроидам кластеров, произвольные единицы. Мужчины.

Для женских голосов имеет место похожая картина. В связи с этим на Рис. 7 приводятся только те импульсы, которые были получены при кластеризации в метрике L_2 .

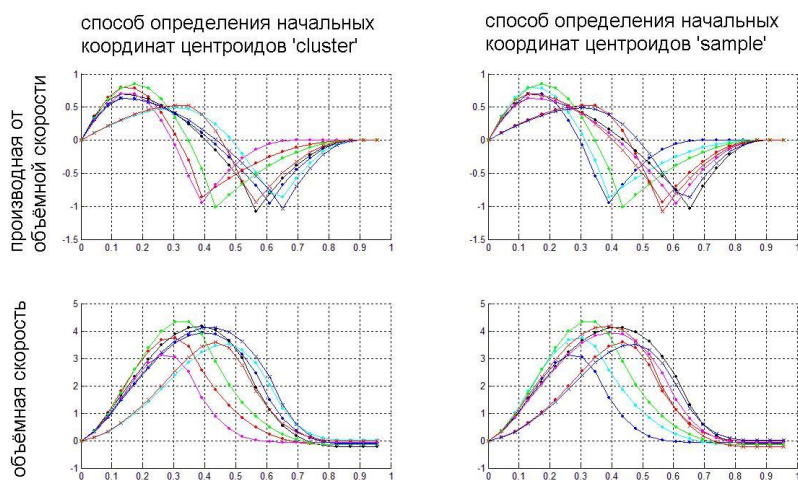


Рис. 7. Импульсы голосового источника и объемной скорости соответствующие центроидам кластеров, метрика L_2 , произвольные единицы. Женщины.

Результаты кластеризации в метрике L_2 представлены в Табл. 2, 3 для мужчин и женщин, соответственно.

Таблица 2. Параметры центроидов. Мужчины.

№ центроида	T1	A2	T2	T3
1	0.237	0.924	0.457	0.881
2	0.253	0.885	0.427	0.855
3	0.356	0.991	0.543	0.824
4	0.193	0.729	0.412	0.830
5	0.204	0.965	0.398	0.666
6	0.119	0.957	0.628	0.858
7	0.329	0.975	0.484	0.696

Таблица 3. Параметры центроидов. Женщины.

№ центроида	T1	A2	T2	T3
1	0.156	0.982	0.389	0.679
2	0.331	0.989	0.560	0.887
3	0.130	0.984	0.662	0.897
4	0.148	0.832	0.403	0.871
5	0.139	0.931	0.616	0.892
6	0.174	0.976	0.446	0.894
7	0.316	0.981	0.639	0.895
8	0.132	0.974	0.585	0.892

Некоторой количественной мерой свойств найденного множества центроидов является матрица их взаимных расстояний в L_2 (Табл. 4, 5).

Таблица 4. Матрица расстояний между центроидами. Мужчины.

Номер вектора	1	2	3	4	5	6	7
1	0	0.3239	0.2890	0.3194	0.3353	0.2111	0.1698
2		0	0.3114	0.3020	0.2558	0.2117	0.2517
3			0	0.1547	0.2637	0.2287	0.2136
4				0	0.1442	0.2137	0.2059
5					0	0.1709	0.1905
6						0	0.0571
7							0

Из Табл. 4 видно, что расстояние между шестым и седьмым векторами значительно меньше расстояний между остальными парами векторов (0.0571). Поэтому, при необходимости уменьшить объем кодовой книги с целью ускорения, можно оставить только первые 6 векторов начального приближения.

Число векторов начальных приближений, найденных путем кластеризации, для женщин равно 8. Их взаимные расстояния показаны в Табл. 5.

Таблица 5. Матрица расстояний между центроидами. Женщины.

Номер вектора	1	2	3	4	5	6	7	8
1	0	0.3207	0.3510	0.2445	0.3164	0.2235	0.3676	0.2910
2		0	0.2258	0.2882	0.2088	0.1946	0.0823	0.2014
3			0	0.3024	0.0706	0.2207	0.1868	0.0779
4				0	0.2366	0.1542	0.3272	0.2329
5					0	0.1796	0.1854	0.0536
6						0	0.2401	0.1454
7							0	0.1918
8								0

Здесь также имеются пары близких векторов: это третий – пятый, второй – седьмой, третий – восьмой и пятый – восьмой векторы. Анализируя эту матрицу, при необходимости экономии объема кодовой книги, можно оставить только пять векторов: первый, третий, четвертый, шестой и седьмой.

Результаты вычеркивания векторов можно оценить на Рис. 8.

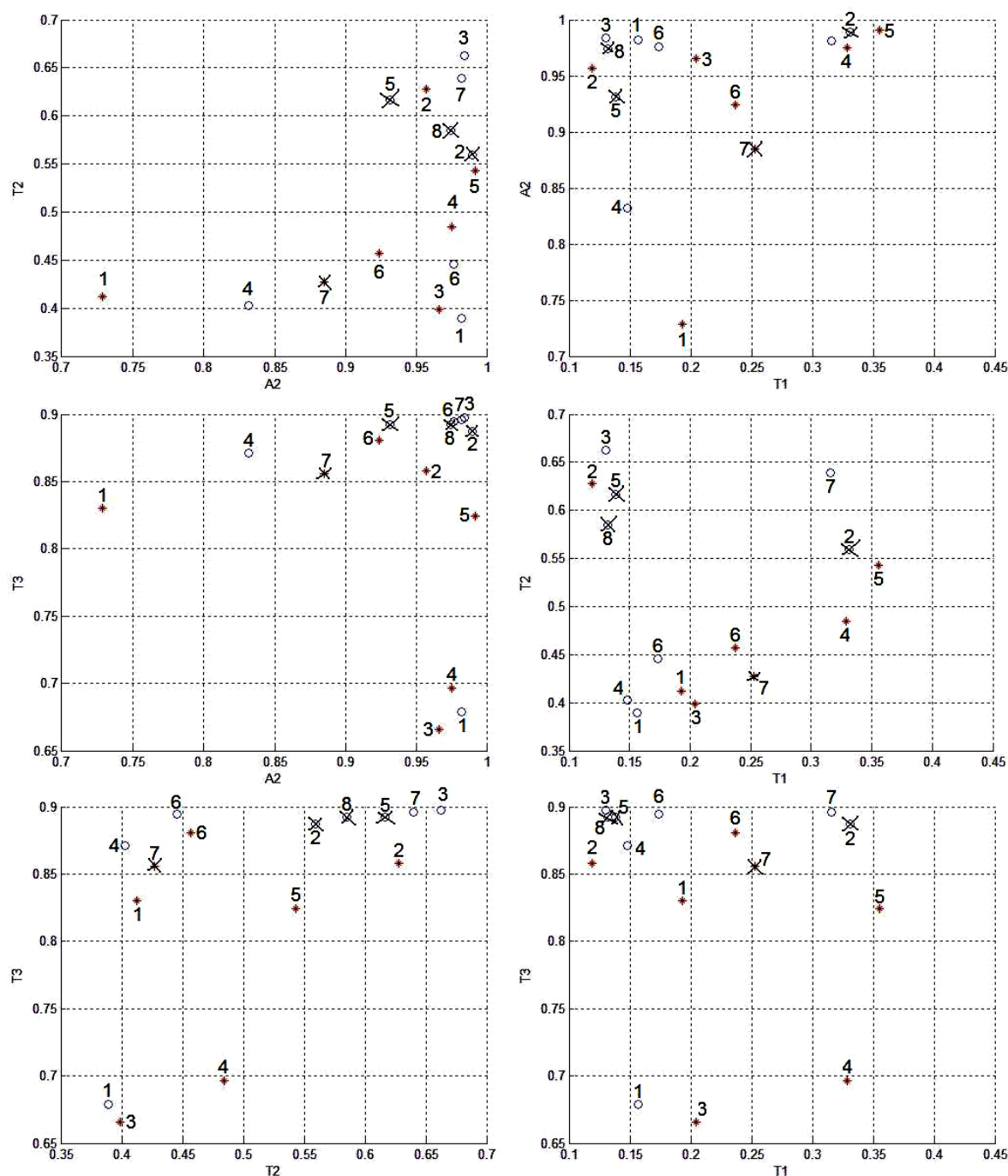


Рисунок 8. Проекция векторов начальных приближений для мужчин (*) и для женщин (o). Номера точек совпадают с номерами векторов по таблицам 2 и 3. Перечёркнутые значки - это отброшенные векторы.

На Рис. 8 видно, что вектора начальных приближений соответствующие мужчинам, даже приблизительно, не имеют ничего общего с векторами для женщин. Поэтому в задачах распознавания речи независимо от диктора или идентификации диктора при определении параметров голосового источника нельзя обойтись без предварительно распознавания пола говорящего. С другой стороны, начальные вектора оказались практически независимыми от частоты основного тона. В этом нет противоречия, поскольку временные параметры модели голосового источника нормированы к периоду основного тона.

Оценка чувствительности найденных параметров к индивидуальным особенностям голоса дикторов выполнялась для параметров модели функции площади сечения голосовой щели $S_m(t)$, предложенной в [1]:

$$S_m(t) = \begin{cases} S_{\max} \left[\sin \left(\frac{\pi t}{2t_1 T_0} \right) \right]^p, & 0 \leq t \leq t_1 T_0 \\ S_{\max} \left[\cos \left(\frac{\pi(t - t_1 T_0)}{2(t_2 - t_1) T_0} \right) \right]^q, & t_1 T_0 < t \leq t_2 T_0 \\ 0, & t_2 T_0 < t \leq T_0 \end{cases} \quad (9)$$

Здесь параметры t_1 и t_2 соответствуют моментам времени, когда достигается максимальное значение площади и полное закрытие голосовой щели. Параметры p и q характеризуют скорость нарастания или спада. Функция $S_m(t)$ вычисляется по найденной объемной скорости потока через голосовую щель [6].

Мера чувствительности параметров определяется как площадь под плотностью распределения параметра P_1 для данного диктора и суммарной плотностью вероятности P_2 для всех остальных дикторов в базе данных в области их пересечения. На Рис. 9 эта область заштрихована на интервале $[a, b]$, а координатные оси представлены в произвольных единицах. Выбор такой меры определяется алгоритмом принятия решений, по которому неизвестный параметр считается принадлежащим к первому распределению, если $P_1/P_2 \geq 1$, и принадлежащим второму распределению, если $P_1/P_2 < 1$. Порог принятия решения равен положению точки c на оси абсцисс.

Площадь под распределениями в области пересечения интерпретируется как суммарная ошибка ε того, что реализация, в действительности принадлежащая первому распределению, будет распознана как принадлежащая второму распределению и реализация, в действительности принадлежащая второму распределению, будет распознана как принадлежащая первому распределению.

$$\varepsilon_i = \int_a^b \min(P_1(x_i), P_2(x_i)) dx_i$$

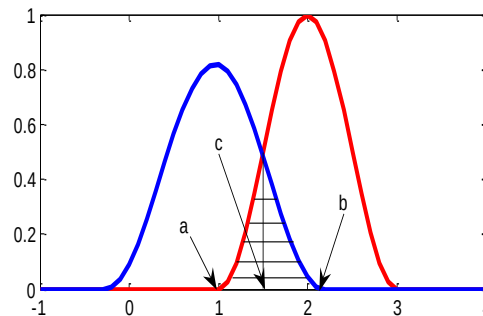


Рис. 9. Иллюстрация оценки суммарной ошибки распознавания.

Ошибки распознавания при оптимизации с найденными начальными приближениями оказались сопоставимы с ошибками, полученными при использовании "нулевых" начальных приближений. При этом, как в первом, так и во втором случае, чувствительность к параметрам женского голоса ниже, чем к параметрам мужского голоса. Это может быть следствием объективной похожести женских голосов, но может также свидетельствовать и о том, что модель голосового источника лучше описывает свойства мужских голосов, чем женских. О такой возможности упоминается в [7], где было установлено, что модель источника, предложенная в [8], описывает форму импульса голосового источника для женских голосов хуже, чем для мужских голосов. Для нормальных голосов средне-квадратическая ошибка может достигать 20%, и погрешность значительно увеличивается даже при слабой патологии голоса.

Таким образом, переход от произвольных начальных приближений к приближениям, найденным как центроиды кластеров, позволил ускорить вычисления за счет сокращения их числа, сохранив при этом чувствительность к индивидуальным особенностям источника голосового возбуждения у разных людей. Это открывает благоприятные возможности для применения разработанного метода в практических приложениях, таких как диагностика патологии гортани и распознавание диктора.

Литература

1. Сорокин В.Н., Макаров И.С. (2008). Распознавание пола диктора по голосу. Акустический ж., т. 54, №4, 1-9.
2. Макаров И.С., Сорокин В.Н. (2008). Распознавание диктора по параметрам голосового источника. Труды XX сессии Российского акустического общества, т. 3, 48 – 51.
3. Маркел Д., Грей А. (1980). *Линейное предсказание речи*. - М.: Связь. – 308 с.
4. Сорокин В.Н., Макаров И.С. (2006). Обратная задача для голосового источника. Информационные процессы. т.6. №4, 375-395.
5. Цыплихин А.И. (2007). Анализ импульсов голосового источника. Акустический журнал.т. 53. № 1, 119-133.
6. Сорокин В.Н. (1992). *Синтез речи*. –М.: Наука, 392 с.
7. Arroabarren I., Carlosena A. (2003). Glottal source parameterization: a comparative study. Proc. VOQUAL'03, 29-34.
8. Fant G., Liljencrants J., Lin Q. (1985). A Four Parameter Model of Glottal Flow. STL-QPSR, v. 4, 1-13.