Устойчивость оценок формантных частот

В.Н. Сорокин
доктор физико-математических наук

А.С. Леонов,
доктор физико-математических наук

И.С. Макаров

Выполнен сравнительный анализ точности и устойчивости мгновенных оценок формантных частот в речевом сегменте методом нулей сигнала и различными модификациями метода линейного предсказания для синтезированных звуков и сигналов, параллельно записанных с микрофонов разного типа. Все использованные методы линейного предсказания показали существенно больший разброс оценок, чем метод нулей сигнала. Установлено, что стабилизация мгновенных оценок формантных частот достигается путём использования информации о характерных акустических характеристиках гласноподобных звуков в конкретном языке. Устойчивость определения формантных треков обеспечивается путём их аппроксимации кусочно-линейными функциями.

Введение

Для решения обратной задачи нахождения формы речевого тракта по сегменту речи нужно оценить резонансные частоты тракта, используя речевой сигнал. Он, однако, определяется не только резонансами речевого тракта, но и резонансами подсвязочной области — трахеи, бронхов и лёгких. Кроме того, в нём присутствуют резонансы носовой полости, причём не только для назальных согласных, но и для назализованных гласных. Поэтому выбор резонансных частот, принадлежащих только речевому тракту выше голосовой щели, представляет собой трудную задачу. Более того, оценка формантных частот тракта по сигналу есть некорректно поставленная задача, что может выражаться в неоднозначности решения (при наличии близких резонансов) и его неустойчивости по отношению к погрешностям измерений. Последние связаны с искажениями сигнала каналом регистрации, реверберацией помещения, нестабильностью частоты основного тона и другими факторами. Амплитудные и частотные модуляции формант усугубляют неоднозначность оценок их частоты.

Оценки формантных частот выполняют как в частотной, так и во временной области. Один из самых распространённых подходов основан на ***методах линейного предсказания***, которые предназначены для описания сигнала во временной области. В целом, эти методы могут давать удовлетворительные оценки формант. Однако многолетние исследования такого подхода показали, что любые модификации методов линейного предсказания неустойчивы относительно аддитивных шумов, особенно при оценке низкочастотных формант. Даже при относительно хороших условиях измерений погрешность оценки формант методами линейного предсказания, как правило, не ниже 10% и к тому же зависит от частоты основного тона [1].

Метод нулей сигнала для оценки формантных частот [5, 6] основан на анализе распределения длительностей интервалов между нулями сигнала. Идеи, лежащие в основе метода, использованы ещё в первых работах по исследованию проблемы автоматического распознавания речи. В своих первых реализациях (на аналоговых устройствах) метод обычно применялся к так называемому клиппированному сигналу. Последний получался путём использования очень большого коэффициента усиления с последующим ограничением амплитуды. В результате преобразованный речевой сигнал представлялся в виде последовательности прямоугольных импульсов с фиксированной амплитудой [2]. Это было удобно для обработки сигнала аналоговой аппаратурой. Но оказалось, что клиппированный сигнал имеет низкую помехоустойчивость, и в результате метод оценки формант с помощью выделения нулей сигнала был на какое-то время забыт.

Развитие цифровой техники привело к возрождению интереса к методу нулей сигнала. В своём новом варианте [5,6] метод оказался более помехоустойчивым, чем методы линейного предсказания и спектрального анализа. Кроме того, метод нулей позволяет обнаружить тонкую структуру динамики формант [5,6]. В работах [3,4] показано, что один из вариантов метода нулей, рассмотренный там под названием «zero-crossing», превосходит известные методы линейного предсказания для низких формант вплоть до SNR=0 dB.

1. Алгоритм метода нулей сигнала

Особенность метода нулей заключается в игнорировании формы колебаний. При этом, конечно, теряется часть информации. Поэтому при низком уровне шумов такие методы, как автокорреляционный или линейное предсказание, могут иметь преимущество. Однако форма колебаний искажается по мере роста уровня шумов, и это преимущество превращается в недостаток.

В большинстве методов оценки формант применяется предварительная обработка сигнала с помощью набора пересекающихся полосовых фильтров. Тип фильтров, их полоса и степень перекрытия влияют на качество последующего анализа и итоговых оценок резонансных частот. Использование фильтров в полосах, примерно соответствующих диапазонам положения формант, способствует повышению точности и устойчивости оценок. После подобной предварительной обработки иногда применяется адаптивный фильтр для уточнения положения формант формантных частот [7]. Анализ нулей сигнала предполагает, что в данной частотной полосе присутствуют колебания только одной форманты. Это связано с известным свойством,

согласно которому при наличии нескольких частот средняя частота переходов определяется как средневзвешенная по амплитудам каждой частоты. Именно поэтому в методе нулей сигнала особенно важен выбор полос частот для анализа.

В данной работе рассматриваются три метода предварительной фильтрации сигнала в частотных диапазонах, где ожидается присутствие только одной форманты.

В первом методе частотные полосы фильтров устанавливаются следующим образом. Первая форманта любого звука анализируется в двух фильтрах с полосами 130 Гц — 400 Гц (фильтр $Ф_{11}$) и 300 Гц — 800 Гц (фильтр $Ф_{12}$). Второй форманте соответствуют три фильтра: 700 Гц — 1600 Гц (фильтр $Ф_{21}$), 1000 Гц — 2000 Гц (фильтр $Ф_{22}$) и 1400 Гц — 2400 Гц (фильтр $Ф_{23}$). Наконец, третья форманта ожидается в одном из двух фильтров с полосами 1700 Гц — 2500 Гц (фильтр $Ф_{31}$) и 2300 Гц — 3500 Гц (фильтр $Ф_{32}$). Эти фильтры перекрываются, в результате чего в один фильтр могут попасть колебания, отвечающие двум формантам.

Второй метод использует распределения формант для каждого гласного русского языка в предположении, что тип гласного и пол диктора известны. Диапазоны формантных частот для некоторых русских гласных даны в таблицах 1, 2.

Таблица 1

Диапазоны формантных частот гласных русского языка для мужских голосов

Гласный	F1 Гц	F2 Гц	F3 Гц
А	450–850	950–1500	1900–2950
Э	320–530	1450–2250	2000–2950
О	300–750	600–1400	1800–3200
И	200–550	1650–2750	2250–3500
Ы	210–500	1650–2600	2150–3100
Е*	250–570	1450–2550	2150–3350
Я*	330–750	1350–2200	2000–3100

* В позиции между мягкими согласными.

Таблица 2

Диапазоны формантных частот гласных русского языка для женских голосов

Гласный	F1 Гц	F2 Гц	F3 Гц
А	550–1000	1100–1650	1950–3100
Э	350–600	1800–2600	2350–3350
О	320–850	600–1550	1800–3300
И	220–620	1850–3100	2550–3600
Ы	250–580	1900–2950	2300–3600
Е*	300–650	2000–2950	2650–3650
Я*	400–900	1800–2650	2300–3500

* В позиции между мягкими согласными.

Третий метод использует параллельный анализ сигнала в полосах, характерных для всех гласных русского языка, в случае когда тип гласного неизвестен или наблюдается переход от одного гласного к другому. Окончательный выбор оценок формантных частот выполняется по критерию, включающему вероятность попадания в трёхмерный вектор формантных частот и суммарную энергию сигнала на этих частотах. Если неизвестен и пол диктора, то сигнал анализируется в частотных полосах, установленных и для мужчин, и для женщин. При этом может использоваться информация, найденная из анализа формы голосового источника. Как показано в [8], вероятность правильного определения пола диктора составляет около 90%. С теоретической точки зрения, частоты резонансов речевого тракта могут изменяться на периоде основного тона вследствие изменения граничных условий при переходе от открытой голосовой щели к закрытой. Кроме того, частоты формант в речевом сигнале подвержены влиянию голосового источника. Поэтому оценку формантных частот целесообразно выполнять на интервале закрытой голосовой щели. В данной работе этот интервал определяется как область, примерно равная 30% от периода основного тона, смещённая на 1 мс относительно пиков огибающей по Гильберту в каждой частотной полосе. Эти пики соответствуют всплеску энергии колебаний резонанса после смыкания голосовой щели.

Во всех методах после фильтрации исходного сигнала с помощью каждого из используемых фильтров определяется среднее значение разности времени между нулями отфильтрованного сигнала на интервале закрытой голосовой щели. Это значение принимается как оценка полупериода формантного колебания из рассматриваемого частотного диапазона. Если оказывается, что нулей меньше двух, то оценка не производится. Затем находится среднее значение формантной частоты для нескольких периодов основного тона, формируется узкополосный фильтр с центральной частотой, равной этой средней частоте, и после новой фильтрации исходного сигнала уточняются оценки частот колебаний на данном интервале времени.

На этом же интервале времени вычисляется среднее значение энергии колебаний, и в качестве предварительной оценки частоты форманты выбирается оценка из того фильтра, где энергия наибольшая. При этом отсеиваются оценки, выходящие за пределы диапазона, а среди конкурирующих оценок выбирается та, которая ближе к среднему значению диапазона.

2. Сравнительное тестирование методов формантного анализа

Ниже приводятся результаты сравнения оценок формант методами типа линейного предсказания и методом нулей сигнала. Изучалась точность и устойчивость методов по отношению к аддитивным помехам, типу микрофона и реверберации помещения, а также устойчивость оценки формантных частот в естественной речи.

2.1. Устойчивость относительно аддитивного белого шума

Погрешность определения формантных частот обычно оценивается в экспериментах с синтезированными звуками, поскольку нет другого способа полу-

чить сигнал с известными параметрами. Однако этому методу присущи недостатки, которые не позволяют безоговорочно опираться на результаты такого тестирования. Синтетический сигнал — это суперпозиция колебаний нескольких осцилляторов с собственными частотами, близкими к реальным формантным частотам, под воздействием источника возбуждения, который по своим характеристикам близок к реальному голосовому источнику. Возбуждаемые этим источником парциальные колебания отличаются от собственных колебаний осцилляторов даже на временных участках, соответствующих закрытой голосовой щели. Это отличие вносит ошибку в оценки собственных частот.

Сигнал, синтезированный с помощью суммирования экспоненциально затухающих колебаний, должен был бы наилучшим образом соответствовать анализу методом линейного предсказания, где используется модель, состоящая из набора полюсов. Поэтому этот метод имеет преимущество перед методами, не опирающимися на такую модель. Тем не менее, в присутствии помех даже для синтезированных сигналов метод линейного предсказания не всегда оказывается наилучшим.

Эксперименты по сравнительной оценке точности и устойчивости методов анализа частотного состава синтетических гласных /*А, И, У*/ проводились при наличии помех типа белого шума разного уровня с гауссовым распределением. Для каждого уровня шумов проводилось по 100 испытаний. Использовались разные варианты линейного предсказания: автокорреляционный, ковариационный, метод усечённого сингулярного разложения матриц, метод регуляризации по Тихонову и метод DAP, разработанный в [9] специально для повышения точности анализа женских голосов.

Оценки формантных частот, полученные с помощью линейного предсказания в кратковременном окне анализа, подвергались сортировке. Из множества исходных оценок удалялись действительные полюсы, а также комплексно-сопряжённые полюсы, частота которых ниже некоторого порога (например, 200 Гц). Кроме того, удалялись полюсы, ширина которых превышает некоторый порог (например, 500 Гц).

Результаты анализа этими методами и разработанным нами методом нулей сигнала показаны в таблицах 3, 4 и 5.

Таблица 3

Относительные ошибки (в %) вычисления формантных частот при отношении сигнал/шум SNR = 20 dB

	Autocorrelation LPC			Covariance LPC			DAP			Метод нулей		
	dF1	dF2	dF3	dF1	dF2	dF3	dF1	dF2	dF3	dF1	dF2	dF3
A	-0.8	-0.2	-0.6	-0.3	-0.3	-0.4	-0.1	-0.1	-0.7	-3.6	-1.2	2.0
I	5.8	-1.1	0.3	4.3	-1.8	0.3	2.5	-1.1	0.1	-5.7	-2.8	-0.4
U	11.5	7.2	-5.9	11.1	6.7	-6.0	6.4	5.6	-6.0	-11.1	2.7	2.8

Таблица 4

Относительные ошибки (в %) вычисления формантных частот при отношении сигнал/шум SNR = 15 dB

	Autocorrelation LPC			Covariance LPC			DAP			Метод нулей		
	dF1	dF2	dF3	dF1	dF2	dF3	dF1	dF2	dF3	dF1	dF2	dF3
A	-1.7	-0.5	-0.5	-0.9	-0.6	-0.6	-0.2	-0.4	-0.4	-3.7	-2.2	3.2
I	7.5	0.7	0.6	6.2	0.9	0.6	3.7	0.4	0.4	-5.8	-7.4	-0.4
U	26.9	17.7	-5.9	26.8	17.6	-5.8	18.5	13.0	-6.1	-10.9	6.0	2.4

Таблица 5

Относительные ошибки (в %) вычисления формантных частот при отношении сигнал/шум SN = 10 dB

	Autocorrelation LPC			Covariance LPC			DAP			Метод нулей		
	dF1	dF2	dF3	dF1	dF2	dF3	dF1	dF2	dF3	dF1	dF2	dF3
A	-1.6	-0.0	-1.6	-0.6	-0.0	-0.6	-2.0	-0.2	-0.1	-3.9	-1.7	4.3
I	11.2	0.9	0.8	10.2	1.0	0.8	6.2	0.7	0.7	-5.6	-13.3	3.2
U	36.2	35.0	-3.0	36.0	35.0	-3.1	31.1	28.3	-3.0	-11.1	15.8	4.8

Таблицы подтверждают установленное другими исследователями свойство неустойчивости к шумам оценок формант методами линейного предсказания, особенно заметное при оценке низких частот. Оценка методом нулей сигнала несколько уступает по точности методам линейного предсказания при низком уровне шума, но оказывается значительно устойчивее при высоком уровне шума. Этот же вывод действителен и для других испытанных нами методов линейного предсказания, не показанных в таблицах 3–5.

2.2. Устойчивость относительно типа микрофона

Сравнение точности и устойчивости методов определения формантных частот с использованием только синтетических звуков не гарантирует полной объективности. Именно поэтому мы провели сравнение результатов оценки формант одного и того же речевого сигнала, записанного одновременно разными приемниками звука. Разница в вычисленных формантах характеризует устойчивость метода относительно искажений амплитудно-частотной характеристики канала связи.

Эксперименты по сравнению устойчивости методов относительно типа микрофона выполнялись на речевых сигналах, отобранных из базы данных для русских числительных. В первой группе дикторов речевые сигналы записывались параллельно через телефонную трубку в стандартном положении и в направленный микрофон, укреплённый вертикально на груди диктора. Во второй группе дикторов использовалась телефонная трубка другого типа и кардиоидный микрофон на груди диктора. В третьей группе дикторов

речевой сигнал записывался через микрофон на головной гарнитуре и кардиоидный микрофон, установленный на мониторе компьютера на расстоянии примерно 50–70 см от диктора. Для экспериментов были случайно отобраны по одному мужчине и одной женщине из каждой группы. Из речевых сегментов каждого числительного были вырезаны стационарные участки ударных гласных, которые и подвергались анализу.

Результаты сравнения для метода нулей сигнала и метода DAP приведены в таблицах 6–9.

Таблица 6

Расхождение оценок формантных частот (%). Метод нулей сигнала. Мужчины

	Первая группа			Вторая группа			Третья группа		
Гласный	**dF1**	**dF2**	**dF3**	**dF1**	**dF2**	**dF3**	**dF1**	**dF2**	**dF3**
нОль	0.0760	0.0578	0.0972	0.0324	0.0216	0.0975	**0.2002**	0.0449	0.0273
одИн	0.0468	0.0221	0.0918	**0.1682**	0.0069	0.0081	0.0783	0.0283	0.0010
двА	0.0096	**0.1015**	0.0037	0.0246	0.0295	**0.1320**	0.0321	0.0062	0.0095
трИ	0.0644	0.0488	**0.1312**	**0.1560**	0.0479	0.0548	0.0632	0.0003	0.0953
четЫре	0.0571	0.0386	0.0595	**0.2160**	0.0723	0.0645	**0.1071**	0.0050	**0.1618**
пЯть	0.0091	0.0228	0.0912	0.0272	0.0136	0.0083	0.0011	0.0118	**0.1881**
шЭсть	0.0015	0.0120	0.0568	0.0665	0.0430	**0.1058**	**0.1046**	0.0150	0.0826
сЕмь	0.0060	0.0062	0.0826	0.0661	0.0158	0.0136	**0.2350**	0.0174	0.0556
вОсемь	0.0182	**0.1325**	0.0029	0.0888	0.0739	**0.2224**	**0.1012**	0.0372	0.0742
дЕвять	0.0085	0.0043	0.0295	**0.2536**	0.0435	**0.1017**	**0.1060**	0.0002	**0.1035**
Среднее	0.0297	0.0447	0.0646	0.1099	0.0368	0.0809	**0.1029**	0.0166	0.0799

Среднее — 6.4%

Средняя ошибка оценки формант по методу нулей сигнала для всех гласных у мужчин составляет: в первой группе — 4.6%, во второй группе — 7.6%, а в третьей группе — 6.6%. Количество рассогласований оценок с уровнем от 10% до 20% равно 14, а с уровнем от 20% до 30% равно 5.

Таблица 7

Расхождение оценок формантных частот (%).
Метод линейного предсказания DAP с предыскажением. Мужчины

	Первая группа			Вторая группа			Третья группа		
Гласный	**dF1**	**dF2**	**dF3**	**dF1**	**dF2**	**dF3**	**dF1**	**dF2**	**dF3**
нОль	0.0603	0.0380	**0.1248**	0.0116	0.0649	0.0011	0.0246	0.0718	0.0271
одИн	0.0323	0.0608	0.0640	**0.1633**	0.0206	0.0403	**0.2977**	0.0170	0.0312
двА	0.0009	0.0641	**0.2444**	0.0053	0.0166	0.0137	0.0482	0.0024	0.0276
трИ	**0.1020**	0.0779	0.0085	**0.2095**	**0.1604**	0.0895	0.2298	0.0385	0.0004

четЫре	0.0046	0.0302	0.0507	0.2759	0.0646	0.0170	0.1122	0.0678	0.0166
пЯть	0.0174	0.0224	0.0142	0.0395	0.0120	0.0131	0.0217	0.0080	0.0545
шЭсть	0.0012	0.0014	0.0030	**0.1184**	0.0175	0.0880	NaN	**0.1024**	0.0067
сЕмь	0.0797	0.0023	0.0270	**0.1494**	0.0308	0.0040	0.2221	0.0769	**0.1103**
вОсемь	0.0561	NaN	0.0003	0.0750	0.0269	0.0006	**0.1994**	0.0703	0.0823
дЕвять	**0.1598**	0.0126	0.0051	**0.1986**	**0.1032**	0.0079	**0.1402**	0.0166	0.0379
Среднее	0.0514	0.0344	0.0542	**0.1246**	0.0517	0.0275	**0.1440**	0.0472	0.0395

Среднее — 6.4%

Средняя ошибка оценки формант методом линейного предсказания по всем гласным у мужчин составляет: в первой группе — 4.7%, во второй группе — 7.6%, а в третьей группе — 6.6%. Так же, как и в методе нулей сигнала, количество рассогласований оценок с уровнем от 10% до 20% равно 14, а с уровнем от 20% до 30% равно 5. Без учёта грубых ошибок метода DAP средняя ошибка по всем измерениям в обоих методах одинакова. Однако имеются две грубые ошибки, когда оценка форманты по методу DAP выходит за ожидаемый диапазон значений формант.

Таблица 8

Расхождение оценок формантных частот (%).
Метод нулей сигнала. Женщины

	Первая группа			Вторая группа			Третья группа		
Гласный	**dF1**	**dF2**	**dF3**	**dF1**	**dF2**	**dF3**	**dF1**	**dF2**	**dF3**
нОль	0.0148	**0.1785**	0.0223	0.0654	0.0230	**0.1349**	0.0236	**0.1150**	0.0806
одИн	0.0061	0.0741	0.0691	0.0842	0.0719	0.0170	**0.1565**	0.0196	0.0492
двА	0.0224	0.0139	**0.1023**	0.0112	**0.1048**	**0.1003**	0.0185	0.0035	**0.1011**
трИ	0.0865	0.0559	0.0440	0.0276	**0.1805**	0.0026	**0.1496**	0.0379	0.0319
четЫре	**0.1445**	0.0400	0.0222	**0.1665**	0.0111	0.0057	0.1069	0.0089	0.0305
пЯть	0.0062	0.0033	0.0565	0.0304	0.0254	**0.1200**	0.0253	0.0256	0.0720
шЭсть	0.0712	0.0033	0.0152	0.0719	0.0058	0.0174	0.0246	0.0316	0.0146
сЕмь	0.0592	0.0365	**0.1065**	0.0849	0.0348	0.0238	**0.1494**	0.0214	0.0106
вОсемь	**0.1221**	0.0298	0.0636	0.0662	0.0107	0.2079	**0.1359**	**0.1606**	0.0528
дЕвять	0.0128	0.0591	0.0389	0.0394	0.0217	0.0096	0.0446	0.0792	0.0140
Среднее	0.0546	0.0494	0.0541	0.0648	0.0490	0.0639	0.0835	0.0503	0.0457

Среднее — 5.7%

У женщин средняя ошибка оценки формант по методу нулей сигнала (для всех гласных) составляет: в первой группе — 5.3%, во второй группе — 5.9%, а в третьей группе — 6.0%. Количество рассогласований оценок с уровнем от 10% до 20% равно 18, а с уровнем от 20% до 30% равно 1.

Таблица 9

**Расхождение оценок формантных частот (%).
Метод линейного предсказания DAP с предыскажением. Женщины**

	Первая группа			Вторая группа			Третья группа		
Гласный	**dF1**	**dF2**	**dF3**	**dF1**	**dF2**	**dF3**	**dF1**	**dF2**	**dF3**
нОль	0.0143	0.0304	**0.1704**	**0.2306**	0.0194	**0.1547**	**0.2103**	**0.4365**	**0.1852**
одИн	0.0066	0.0371	0.0940	**0.3084**	**0.1468**	0.0265	**0.2813**	0.0548	0.0280
двА	0.0029	0.0219	0.0389	0.0090	0.0740	0.0089	0.0011	0.0907	**0.1690**
трИ	0.0151	0.0032	0.0932	**0.4084**	**0.1343**	0.0134	**0.1176**	0.0182	0.0304
четЫре	0.0963	0.0277	0.0537	0.0959	**0.2006**	0.0334	0.0629	**0.1047**	0.0094
пЯть	**0.1451**	0.0251	0.0101	0.0232	0.0401	0.0978	**0.2363**	0.0235	0.0480
шЭсть	0.0096	0.0321	0.0326	0.2413	**0.1237**	**0.1753**	0.0155	0.0191	0.0116
сЕмь	0.0253	0.0291	0.0094	0.0954	**0.1056**	0.0784	**0.1309**	0.0072	0.0734
вОсемь	0.0664	0.0622	**0.1141**	**NaN**	**NaN**	**NaN**	**0.1155**	**0.2953**	0.0368
дЕвять	0.0223	0.0752	0.0461	**0.2611**	**0.1907**	0.0160	0.0204	0.0766	0.0324
Среднее	0.0404	0.0344	0.0663	**0.1859**	**0.1150**	0.0672	**0.1192**	**0.1127**	0.0624

Среднее — 8.9%

Средняя ошибка оценки формант методом линейного предсказания по всем гласным у женщин составляет: в первой группе — 4.7%, во второй группе — 12.3%, а в третьей группе — 9.8%. Количество рассогласований оценок с уровнем от 10% до 20% равно 15, с уровнем от 20% до 30% равно 7. Имеется одна ошибка с уровнем от 30% до 40%, и две ошибки превышают 40%. Кроме того, имеются три грубые ошибки.

Анализ выполнялся первым методом, т.е. в усреднённых диапазонах частот для каждой форманты. Однако в силу того что тип гласного известен, окончательный отбор оценок формант производился с учётом характерного диапазона формантных частот и степени близости к характерному среднему значению каждой форманты гласного.

В таблицах использован термин среды МАТЛАБ–NAN (Not a Number). Он означает, что для одного из микрофонов не найдена оценка форманты в заданном диапазоне. В силу ограниченности тестового материала, разницу в долях процентов можно считать малозначимой, тогда как разница в процентах указывает на определённую тенденцию.

При сравнении данных из таблиц 8 и 9 видно, что число грубых ошибок метода DAP, включая выход за диапазон частот и превышение ошибки в 30% равно 6. Средняя ошибка в методе DAP, специально разработанном для улучшения качества анализа женских голосов, в полтора раза больше, чем в методе нулей сигнала.

Первая и вторая группы отличаются, главным образом, вторым микрофоном, поскольку разницу между двумя типами телефонных трубок можно считать малой по сравнению с разницей между направленным и кардиоидным микрофонами. Разница в оценках формантных частот у мужчин составляет около 3% для обоих методов. Даже при ограниченном речевом материале эта разница представляется значимой. У женщин эта

разница в методе нулей сигнала составляет всего 0.6%, тогда как в методе линейного предсказания она достигает почти 8%.

Итак, оба метода чувствительны к типу микрофона, причём у женщин разница между оценками формант по сигналам от направленного и ненаправленного микрофонов особенно велика в методе линейного предсказания.

2.3. Устойчивость относительно реверберации

Акустические характеристики помещения, в котором происходит запись речевого сигнала, влияют на амплитудно-частотные характеристики сигнала. Это было наглядно продемонстрировано в [10].

Данные таблиц 6–9 позволяют качественно оценить влияние реверберации помещения на погрешность методов анализа. Первая и вторая группы тестов выполнялись на относительно близко расположенных микрофонах, тогда как в третьей группе тестов использовались и близко расположенный ко рту микрофон, и микрофон, удалённый на расстояние в несколько десятков сантиметров. При этом во второй и третьей группах тестов один из микрофонов был один и тот же — микрофон кардиоидного типа, расположенный либо на груди диктора, либо на мониторе.

Средние значения ошибок в методе нулей сигнала для близко расположенных микрофонов и удалённого микрофона оказались довольно близки: 6.1% и 6.6% — у мужчин, и 5.6% и 6.0% — у женщин, так что ошибки отличаются на величину около 0.5%. Для метода линейного предсказания эта разница оказалась больше: 5.7% и 7.5% — у мужчин, и 8.5% и 9.8% — у женщин. В этом случае различие оценок для близких и удалённых микрофонов составила 1.8% и 1.3%. Из этого можно заключить, что реверберация помещения больше сказывается на анализе методом линейного предсказания, чем на анализе методом нулей сигнала.

2.4. Устойчивость анализа натуральных звуков

Число полюсов в амплитудно-частотной характеристике речевого сигнала, оцениваемое методом линейного предсказания, связано с частотой дискретизации сигнала. Поэтому в диапазоне частот, характерных для какого-либо звука речи, может оказаться либо избыточное, либо недостаточное количество полюсов. Это вполне закономерно, поскольку метод линейного предсказания изначально предназначен для аппроксимации сигнала, а не для анализа резонансных частот речевого тракта. Поскольку коэффициенты линейного предсказания вычисляются в процедуре, которая минимизирует ошибку аппроксимации спектра, то количество найденных полюсов и их расположение, вообще говоря, произвольны. И хотя в большинстве случаев вычисленные полюса достаточно близки к резонансам речевого тракта, имеется достаточно много ситуаций, в которых появляются грубые ошибки в оценке формантных частот.

Устойчивость оценок формантных частот методом нулей сигнала зависит от параметров полосовых фильтров и от точности определения интервала

сомкнутых голосовых складок. Как следствие, метод нулей сигнала может ненадёжно определять формантные частоты при сближении формант или на переходных процессах.

Если заранее известно, какой тип гласного соответствует рассматриваемому сегменту речи, как это может иметь место при верификации диктора, то целесообразно использовать фильтры, настроенные на конкретный гласный уже на первом этапе анализа. В этом случае метод нулей сигнала демонстрирует наиболее устойчивые оценки формантных частот. Так, в описанных выше экспериментах по сравнению устойчивости оценок для сигналов, записанных параллельно с микрофонов разных типов, средняя ошибка в методе нулей сигнала для мужчин составила около 4%, а для женщин — около 3%, т.е. в 1.5–2 раза меньше, чем при анализе с помощью фильтров, настроенных на усреднённые диапазоны формант. Это сопоставимо с погрешностью оценок, возникающей из-за дискретизации сигнала по времени.

Если дополнительная информация об ожидаемом типе гласного или переходном процессе отсутствует, то ни один из известных методов анализа формантных частот, включая и разработанный нами метод нулей сигнала, не застрахован от грубых ошибок. Поэтому кажется естественным применить параллельный формантный анализ разными методами. Если бы удалось совместить сильные стороны каждого метода и избежать их недостатков путём формирования критерия выбора оценок, то можно было бы надеяться на получение более точных и устойчивых оценок формантных частот.

Один из вариантов подобного параллельного анализа состоит в использовании метода линейного предсказания для предварительной оценки формантных частот в относительно широких диапазонах возможного положения каждой форманты. Эти оценки используются затем для формирования адаптивных фильтров, выходные сигналы которых анализируются методом нулей сигнала. Недостаток такого подхода состоит в риске грубых ошибок линейного предсказания.

Поиск критерия выбора правильного решения при параллельном использовании разных методов формантного анализа требует специального исследования. В качестве альтернативы такому подходу в данной работе применялся только метод пересечений через нуль, но параллельная оценка выполнялась для всего множества фильтров, соответствующих диапазонам формантных частот каждого гласного.

3. Динамика формантных частот

Известно, что мгновенные оценки формантных частот любым методом ненадёжны. Графически это выглядит как разброс точек на формантных треках (см., например, рис. 4). При этом иногда (например, в случае близких формант) трудно определить, какому треку принадлежит какая точка. Поэтому, получив формантные треки на интервалах времени определённой длительности, обычно выполняют коррекцию ошибок кратковременного анализа путём интерполяции треков.

Многие алгоритмы коррекции формантных треков используют предположение об их непрерывности или гладкости, основанное на непрерывности артикуляторных движений. Например, в классическом алгоритме [11] сначала находятся квазистационарные согласованные сегменты речевого сигнала (так называемые опорные точки), на которых оценки формантных частот наиболее надёжны. Затем алгоритм последовательно про-

должает формантные треки между соседними опорными точками, выбирая при переходе от предыдущего к последующему формантному вектору из множества кандидатов тот, который наиболее близок (в евклидовой метрике) к уже оценённому на предыдущем сегменте. В работе [12] построение формантных треков по оценкам линейного предсказания осуществляется с помощью дискретных марковских моделей. В работах [13, 14] искомые векторы формантных частот выбираются из множества кандидатов с помощью процедуры динамического программирования. При этом используется некоторый составной критерий отбора, который включает в себя невязку соседних по времени векторов формантных частот, условие минимума формантных ширин и близость формант к формантным частотам нейтрального гласного.

Однако на участках переходных процессов в речевом сигнале нередко наблюдается нарушение непрерывности формантных треков. Это явление характерно для женских голосов, хотя у мужских голосов оно также иногда наблюдается. В качестве примера рассмотрим рис. 1 и 2, где показаны сонограммы слогов /*УА*/ и /*АУ*/ для женского голоса. Эти сонограммы демонстрируют не только разрывы треков первой и второй форманты, но и разрывы направления движения формант. Отметим, что эти разрывы не

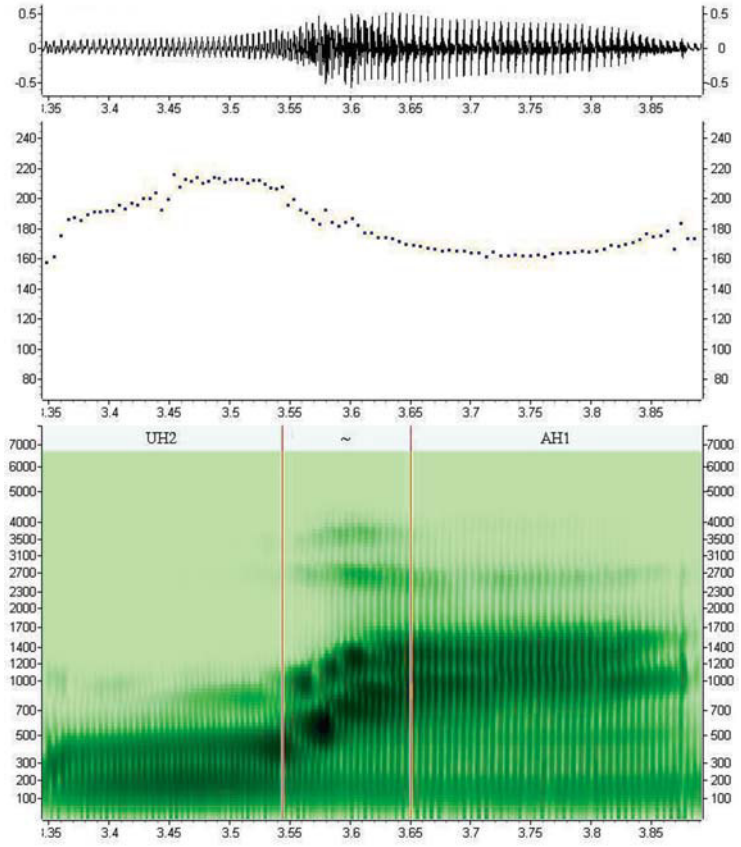

Рис. 1. Звукосочетание /УА/, женский голос.
Вверху – осциллограмма сигнала, в середине – контур основного тона, внизу – сонограмма со шкалой мел по оси частот

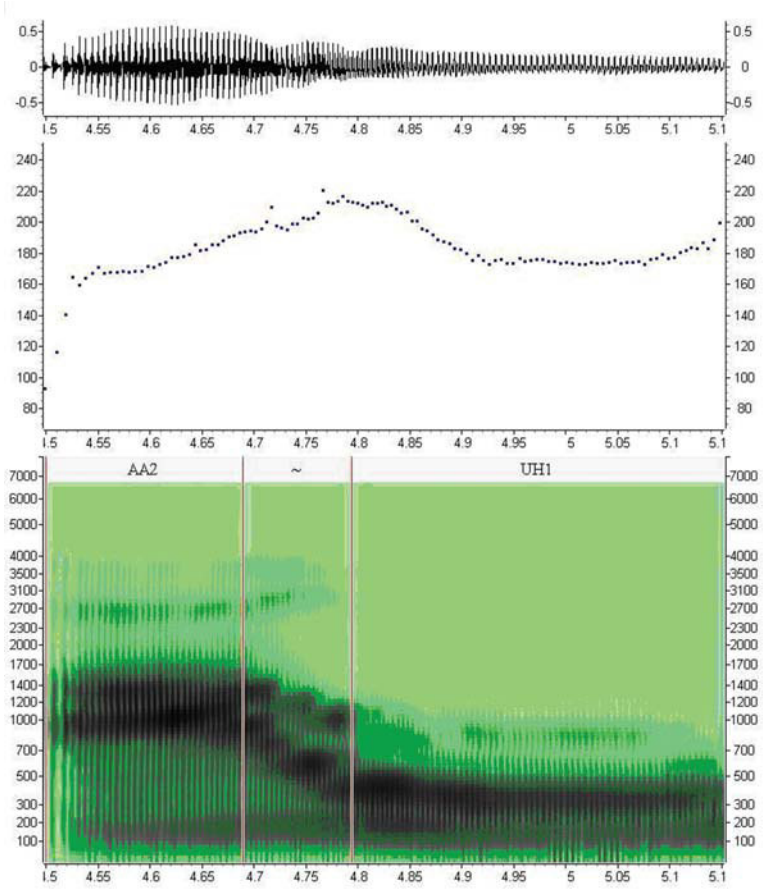

Рис. 2. Звукосочетание /АУ/, женский голос. Вверху – осциллограмма сигнала, в середине – контур основного тона, внизу – сонограмма со шкалой мел по оси частот

регистрируются методами линейного предсказания, и лишь анализ интервалов между нулями сигнала при закрытой голосовой щели обнаруживает эти явления. На обоих рисунках видно, что разрывы в формантных треках сопровождаются амплитудными модуляциями осциллограмм, и даже на квазистационарном участке трека второй форманты наблюдаются довольно длительные спады энергии. Это затрудняет использование информации об амплитуде формант при отслеживании треков оценок формантных частот во времени.

В работе [15] было показано, что разрывы формантных треков в динамическом спектре речевого сигнала могут наблюдаться в тех случаях, когда резонансные частоты речевого тракта и подсвязочной области близки. Близость ротовых и подсвязочных резонансов не является единственной причиной разрывов. Другие факторы и, в частности, соотношение частоты основного тона и частоты форманты, также влияют на форму динамического спектра звуков речи. В особенности это относится к женским голосам с высоким основным тоном. Некоторые математические аспекты этого явления рассмотрены в *Приложении*.

Из сказанного ясно, что коррекцию формантных треков необходимо выполнять, исходя из возможной их разрывности. Соответственно, при интерполяции треков нет оснований для использования непрерывных функций и, в частности, многочленов высоких поряд-

ков. Наиболее целесообразным представляется использование кусочно-линейной аппроксимации треков.

Приведём пример такой коррекции треков формант. На рис. 3 показаны мгновенные оценки формантных треков в слоге /*ИА*/ по методу нулей сигнала и их кусочно-линейная аппроксимация.

На интервале времени вокруг отсчёта 0.25 с наблюдаются скачки всех трёх формант. Особенно велик скачок частоты второй форманты (около 500 Гц). На сонограмме этого слога действительно видны разрывы траекторий формантных частот при переходном процессе от звука /*И*/ к звуку /*А*/. Однако эти скачки заглажены в силу использования весовой функции при вычислении спектра. Лишь мгновенные оценки формантных частот по методу нулей сигнала чётко выявили разрывы формант на переходных участках в звукосочетаниях.

В этом примере исходным материалом для метода нулей являлись отфильтрованные сигналы с фильтрами в характерных диапазонах формантных частот для гласных русского языка. В каждый момент времени параллельно выполнялись оценки по фильтрам, соответствующим формантам гласных /*И*/ и /*А*/. Выбирались оценки того набора фильтров, в котором сумма пиков огибающей по всем трём формантам была наибольшей. Без такого отбора разброс оценок формантных частот слишком велик, и никакое сглаживание не улучшает поведения формантных треков. В частности, если исходить из обычного предположения, что следующее значение частоты некоторой форманты должно находиться как можно ближе к предыдущему, то в области переходного процесса произойдёт перескок оценок второй форманты на третью форманту.

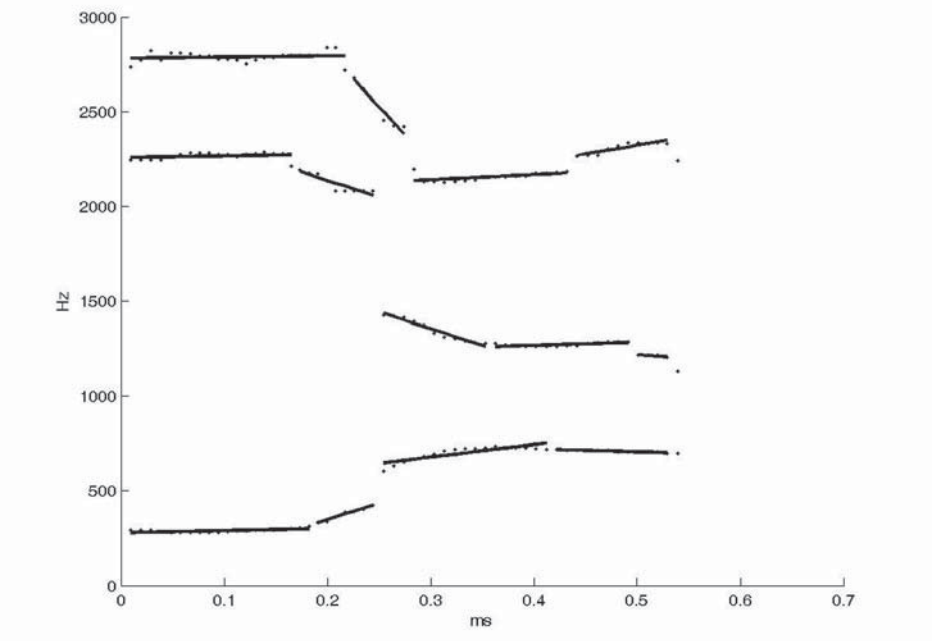

Рис. 3. Мгновенные оценки формантных частот в слоге /ИА/ (···) и кусочно-линейная аппроксимация треков (——)

Успех этого численного эксперимента позволяет сформулировать ещё один способ стабилизации оценок формантных частот, отличный от использования метода линейного предсказания в качестве предварительной оценки. Для каждого языка можно найти небольшое число характерных векторов формантных частот, примерно соответствующих гласным этого языка в том смысле, как их определяют фонетисты. Методика поиска этих характерных векторов путём кластеризации множества измерений формантных частот была описана в [16]. Распределение вероятностей каждого из этих характерных векторов может быть использовано для построения согласованных фильтров. Сигналы на выходе каждого набора фильтров подвергаются анализу согласно некоторому критерию, и оценки формантных частот выбираются для того набора фильтров, где значение этого критерия наилучшее. В частности, этот критерий может состоять в суммарной энергии — так, как это было применено в описанном выше примере.

Очевидно, что такой метод будет лучше всего работать на квазистационарных участках речевого сигнала, тогда как переходные процессы могут оцениваться с большей погрешностью. Однако можно сформировать алгоритм коррекции оценок на переходных процессах, используя устойчивые оценки формантных частот на краях переходного процесса.

Преимущество этого подхода заключается в том, что его можно применять для произвольного контекста, не заботясь о предварительной оценке положения во времени гласноподобных сегментов. При этом полностью используется информация о формантных образах гласных звуков в конкретном языке. Как было показано в данной работе и в предыдущих исследованиях на эту тему [5, 6], без учёта этой информации невозможно сколько-нибудь устойчивое определение формантных частот в речевом сигнале. Ещё одно преимущество заключается в подавлении колебаний, проникающих из подсвязочной области в речевой тракт. Это особенно важно при решении обратной задачи с целью определения формы речевого тракта, для чего нужно быть уверенным в том, что измеренные частоты действительно соответствуют резонансным частотам речевого тракта.

Заметим, что при таком подходе формантный анализ речевого сигнала становится зависимым от конкретного языка, его артикуляторного строя и формантных образов основных гласных. Интуитивно это представляется вполне оправданным. Это также объясняет неудачу многочисленных попыток построить универсальный устойчивый алгоритм определения формантных частот в речевом сигнале независимо от языка. Отсюда можно предположить, что и автоматический анализ взрывных согласных, назальных и фрикативных звуков также должен производиться с использованием специфических акустических свойств конкретного языка. Ясно, что основная трудность при этом заключается в создании достоверной базы акустических характеристик каждого языка на основе более или менее абстрактных методов анализа и ручной обработке полученных данных.

Заключение

Метод нулей сигнала характеризуется значительно меньшим разбросом оценок формантных частот в зависимости от типа регистрирующего микрофона и устойчив к шумам, особенно в низкочастотной области. Мгновенные оценки формантных частот этим методом на периоде основного тона могут быть уточнены (скорректированы) путём использования информации о типе гласного. Предположение о непрерывности формантных

треков при коррекции не оправдано. Поэтому коррекцию оценок формант следует выполнять путём кусочно-линейной аппроксимации с возможными разрывами треков.

ПРИЛОЖЕНИЕ

Экспериментально было установлено, что на оценку формантных частот влияет частота основного тона, причём это влияние особенно заметно сказывается при определении низкочастотных формант. Механизм этого явления был не вполне ясен. Ниже приводятся две простые математические модели, позволяющие изучить воздействие источника возбуждения на спектр сигнала и качественно описать соответствующие эффекты, которые проявляются при формантном анализе.

Представим речевой тракт как совокупность не связанных осцилляторов с собственными частотами *F*, определяющими форманты. Будем сначала считать, что эти осцилляторы колеблются без затухания под действием гармонического источника возбуждения с частотой основного тона F_0. Математически запишем это в виде задачи Коши для колебания *y*(*t*):

$$y''+\omega^2 y = A\sin\Omega t,\ \ y(0)=y'(0)=0\,.$$

Здесь $\omega = 2\pi F$ — собственная (круговая) частота осциллятора, а $\Omega = 2\pi F_0$. Нетрудно видеть, что

$$y(t) = -\frac{A}{\omega^2-\Omega^2}\left(\frac{\Omega}{\omega}\sin\omega t - \sin\Omega t\right).$$

Преобразуем найденное решение, вводя величину $\varphi(t)\in[0,2\pi)$ — решение системы уравнений

$$\cos\varphi(t) = \frac{\Omega}{\omega}\sin(\omega+\Omega)t,\ \ \sin\varphi(t) = 1+\frac{\Omega}{\omega}\cos(\omega+\Omega)t\,,$$

а также числа

$$D = -\frac{A}{\omega^2-\Omega^2}\sqrt{1+\left(\frac{\Omega}{\omega}\right)^2}\ ,\ \ \mu = 2\frac{\Omega}{\omega}\left[1+\left(\frac{\Omega}{\omega}\right)^2\right]^{-1}.$$

В итоге оказывается, что

$$y(t) = D\left[1+\mu\cos(\omega+\Omega)t\right]^{1/2}\cos\left[\Omega t+\varphi(t)\right].$$

Это решение легко интерпретируется при $\mu \ll 1$, то есть для формантных частот, много больших частоты основного тона. В этом случае

$$\varphi(t)\approx\frac{\pi}{2},\ D\approx-\frac{A}{\omega^2}\ \text{и}$$

$$y(t)\approx\frac{A}{\omega^2}\sin\Omega t+\frac{A\mu}{2\omega^2}\cos(\omega+\Omega)t\cdot\sin\Omega t\,, \qquad (1)$$

так что движения осцилляторов представляют собой колебания основного тона, на которые наложены колебания со сдвинутой формантной частотой, промодулированные колебаниями с частотой основного тона. Таким обра-

зом, рассмотренная простейшая модель качественно предсказывает не только модуляцию амплитуд формант из-за воздействия голосового источника возбуждения, но и их сдвиг в сторону увеличения частоты.

Влияние гармоник основного тона на спектр колебаний в речевом тракте можно качественно изучить и для источника более общего вида с учётом затухания собственных колебаний. Полагая, что голосовой источник $f(t)$ — это кусочно-гладкая периодическая функция, разложим её в ряд Фурье на периоде колебаний (например, в ряд Фурье по синусам, если $f(0) = 0$). Тогда задачу определения вынужденных колебаний осциллятора можно записать в виде

$$y'' + 2\gamma\omega y' + \omega^2 y = f(t) = \sum_{n=1}^{\infty} b_n \sin n\Omega t \quad (0 < \gamma < 1), \qquad (2)$$
$$y(0) = y'(0) = 0.$$

Её решение представимо как $y(t) = \sum_{n=1}^{\infty} y_n(t)$, где слагаемые находятся из задачи Коши:

$$y_n'' + 2\gamma\omega y_n' + \omega^2 y_n = b_n \sin n\Omega t, \quad y_n(0) = 0, \; y_n'(0) = 0.$$

Можно вычислить, что $y_n(t) = y_n^{(0)}(t) + y_n^{(1)}(t)$, где

$$y_n^{(0)}(t) = b_n \left[\frac{e^{-\gamma\omega t} \sin\left(\sqrt{1-\gamma^2}\,\omega t\right) n\Omega\left(\Omega^2 + 2\gamma^2\omega^2 - \omega^2\right)}{\sqrt{1-\gamma^2}\,\omega\left(4\omega^2\Omega^2\gamma^2 + (\Omega^2 - \omega^2)^2\right)} + \frac{2e^{-\gamma\omega t}\cos\left(\sqrt{1-\gamma^2}\,\omega t\right)\gamma\omega\Omega}{\left(4\omega^2\Omega^2\gamma^2 + (\Omega^2 - \omega^2)^2\right)} \right]$$

$$y_n^{(1)}(t) = b_n \frac{(\omega^2 - n^2\Omega^2)\sin n\Omega t - 2\gamma n\omega\Omega \cos n\Omega t}{(\omega^2 - n^2\Omega^2)^2 + 4\gamma^2\omega^2 n^2\Omega^2}.$$

Решение задачи (2) $y(t) = \sum_{n=1}^{\infty} y_n(t) = \sum_{n=1}^{\infty} y_n^{(0)}(t) + \sum_{n=1}^{\infty} y_n^{(1)}(t) \equiv y^{(0)}(t) + y^{(1)}(t)$

интерпретируется так: в голосовом тракте существуют не только затухающие собственные колебания $y^{(0)}(t)$, но и колебания $y^{(1)}(t)$, которые определяются частотой основного тона. Это верно даже на интервале закрытой голосовой щели, т.е. для временных интервалов, где $f(t) = 0$. Поэтому при формантном анализе на интервале закрытой голосовой щели на получаемый результат влияет член сигнала

$$y^{(1)}(t) = \sum_{n=1}^{\infty} b_n \frac{(\omega^2 - n^2\Omega^2)\sin n\Omega t - 2\gamma n\omega\Omega \cos n\Omega t}{(\omega^2 - n^2\Omega^2)^2 + 4\gamma^2\omega^2 n^2\Omega^2} =$$

$$= \frac{1}{\omega^2} \sum_{n=1}^{\infty} b_n \frac{(1 - n^2\beta^2)\sin n\Omega t - 2\gamma n\beta \cos n\Omega t}{(1 - n^2\beta^2)^2 + 4\gamma^2 n^2\beta^2} =$$

$$= \frac{1}{\omega^2} \sum_{n=1}^{\infty} b_n \sin(n\Omega t + \varphi_n), \qquad (3)$$

где φ_n есть главное решение системы уравнений

$$\cos\varphi_n = \frac{(1-n^2\beta^2)}{(1-n^2\beta^2)^2+4\gamma^2n^2\beta^2},\ \sin\varphi_n = -\frac{2\gamma n\beta}{(1-n^2\beta^2)^2+4\gamma^2n^2\beta^2},\quad \beta=\frac{\Omega}{\omega}.$$

Слагаемое (3) искажает спектр собственных частот сигнала, в котором в итоге появляются колебания с частотами $n\Omega$. При небольших n частоты $n\Omega$ могут быть сравнимы с формантными. Амплитуды Фурье-гармоник функции (3) суть изменённые в $1/\omega^2$ раз амплитуды гармоник источника $f(t)$. Поэтому искажения оценок формантного анализа наиболее существенны для низких частот, когда отношение $1/\omega^2$ велико. Ещё раз подчеркнём, что сделанные выводы справедливы для любой кусочно-гладкой формы источника возбуждения.

Проведённый анализ объясняет причины экспериментально установленной зависимости оценок формантных частот от частоты основного тона источника голосового возбуждения, которая наблюдается для любых методов формантного анализа.

Литература

1. G.K. Vallabha, B.Tuller (2002). Systematic errors in formant analysis of steady-state vowels. Speech Communication, v.38, pp.141–160.

2. Цемель Г.И. Опознавание речевых сигналов. М.: Наука, 1971.

3. R.J. Niederjohn, M.Lahat (1985). A zero-crossing consistency method for formant tracking of voiced speech in high noise levels. IEEE on Acoustics, Speech and Signal Processing, ASSP–33, N2, 349–355.

4. Th.Sreenivas, R.J. Niederjohn (1992). Zero-crossing based spectral analysis and SVD spectral analysis for formant frequency estimation in noise, IEEE transactions on Signal Processing, v.40, N2, 282–293.

5. Сорокин В.Н., Трифоненков И.П. Об автокорреляционном анализе речевых сигналов. 1996. Акуст. ж., Т. 42. №3. С. 368–374.

6. Леонов А.С., Сорокин В.Н. К анализу резонансных частот речевого тракта. Информационные процессы, Т. 7. 2007. №4, 386–400. www.jip.ru.

7. K.Mystafa, I.C. Bruce (2006). Robust formant tracking for continuous speech with speaker variability. IEEE transactions on Audio, Speech, and Language Processing, v.14, N2, 435–444.

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

9. A.El-Jaroudi, J.Makhoul (1991). Discrete All-Pole Modeling. IEEE Trans. Signal Process., vol.39, No.2, pp.411–423.

10. Сорокин В.Н., Макаров И.С. Обратная задача для голосового источника. Информационные процессы. 2006. Т. 6, №4, 375–395. www.jip.ru.

11. S.McCandless. An Algorithm for Automatic Formant Extraction Using Linear Prediction Spectra. // IEEE Trans. Acoust., Speech, Signal Process., vol.ASSP–22, 1974, pp.135–141.

12. G.Kopec. Formant Tracking Using Hidden Markov Models and Vector Quantization. // IEEE Trans. Acoust., Speech, Signal Process., vol.ASSP–34, 1986, pp.709–729.

13. D.Talkin. Speech Formant Trajectory Estimation Using Dynamic Programming with Modulated Transition Costs. // J.Acoust. Soc. Amer. S1, 1987, p.S55.

14. K.Xia, C.Espy-Wilson. A New Strategy of Formant Tracking Based on Dynamic Programming. // Proc. Int. Conf. Spoken Lang. Process., 2000, pp.55–58.

15. X.Chi, M.Sonderegger (2007). Subglottal coupling and its influence on vowel formants. Journal. Acoust. Soc. Am., v.122, N3, 1735–1745.

16. Сорокин В.Н., Цыплихин А.И. Сегментация и распознавание гласных. Информационные процессы, 2004. Т. 4. №2, С. 202–220.www.jip.ru.

В.Н. Сорокин,
доктор физико-математических наук,
ведущий научный сотрудник
Института проблем передачи информации РАН.
E-mail: vns@iitp.ru.

А.С. Леонов,
доктор физико-математических наук,
профессор кафедры математики
Московского инженерно-физического института
(Федеральный исследовательский ядерный университет).
Специалист в области решения обратных и некорректно
поставленных задач науки и техники
(обратные задачи теплопроводности и диффузии,
задачи обработки изображений,
задачи оптимального синтеза технических систем,
обратные задачи речевых технологий и др.).
Автор монографий по решению нелинейных некорректных задач.

И.С. Макаров,
Институт проблем передачи информации,
Российская академия наук, Москва, Россия.