АКУСТИЧЕСКИЙ ЖУРНАЛ, 2024, том 70, № 5, с. 778–794

ОБРАБОТКА АКУСТИЧЕСКИХ СИГНАЛОВ. КОМПЬЮТЕРНОЕ МОДЕЛИРОВАНИЕ

УДК 612.85

ИМПУЛЬСНЫЙ ИСТОЧНИК ВОЗБУЖДЕНИЯ В РЕЧЕВОМ СИГНАЛЕ

© 2024 г. В. Н. Сорокин

Институт проблем передачи информации РАН, Б. Каретный пер. 19, Москва, 127994 Россия

e-mail: vns@iitp.ru

Поступила в редакцию 18.03.2024 г.
После доработки 12.07.2024 г.
Принята к публикации 05.09.2024 г.

Свойства речевого взрыва смычки исследуются на материале базы данных 39 дикторов, содержащей однозначные и многозначные числительные с параллельной записью сигналов на телефонную трубку и направленный микрофон. Детектирование речевого взрыва выполняется кратковременным и долговременным детектором спектрально-временных неоднородностей, а также детектором меры сходства собственных функций спектра взрыва согласных и текущего спектра речевого взрыва. Вероятность присутствия звонкой или глухой смычки оценивается в пространствах амплитудного спектра и спектра групповой задержки по отношению энергии в области высоких и низких частот. Место артикуляции заднеязычного согласного влияет на распределения вероятности длительности интервала между началом речевого взрыва и началом гласного, частоты пика с максимальной амплитудой в высокочастотной области, отношения энергии в области высоких и низких частот спектра речевого взрыва, а также меры сходства собственных функций спектра взрыва согласного и текущего спектра речевого взрыва.

Ключевые слова: распознавание речи, сегментация речевого сигнала, детектор взрыва смычки, распознавание взрывных согласных

DOI: 10.31857/S0320791924050121, **EDN:** XBBEKN

1. ВВЕДЕНИЕ

В процессах речеобразования принимает участие несколько источников возбуждения акустических колебаний — голосовой, поршневой, турбулентные источники и импульсный источник. Голосовой источник пропорционален производной по времени от объемной скорости потока через голосовую щель. Он кодирует признак звонкости, а частота основного тона формирует интонацию. Синхронный с голосовым источником поршневой источник порождается вертикальными колебаниями голосовых складок, он также пропорционален производной по времени от объемной скорости потока вытесняемого воздуха. Источники широкополосного шума создаются турбулентностью потока на выходе из сужения в речевом тракте. Один такой источник присутствует в области над голосовыми складками. Этот источник вместе с поршневым источником влияет на тембр голоса. Турбулентный источник в верхней и передней части речевого тракта формирует фрикативные звуки, и его спектр кодирует место артикуляции. Импульсный источник в акустике речи возникает после раскрытия смычки взрывных согласных. Он также пропорционален производной по времени от объемной скорости воздушного потока через сужение, и ему сопутствует источник турбулентного шума.

Такой импульсный источник порождает кратковременный сегмент, который называется речевым взрывом. Несмотря на малую энергию речевого взрыва, импульсный источник влияет на восприятие места артикуляции взрывных звуков (координату нулевой площади поперечного сечения речевого тракта). Удаление взрыва начальных глухих согласных ухудшает распознаваемость при сопоставлении слов типа */почка, точка, кочка/*. В экспериментах по восприятию речи было установлено, что динамика формантных частот гласных после взрыва смычки обеспечивает лишь 70% распознавания согласных [1]. В [2] было подтверждено, что достаточно высокий уровень идентификации взрывных согласных (около 87%) достигается только при использовании всей доступной информации о спектре, длительности речевого взрыва, и начальных значений формантных частот последующего гласного. В русском языке импульсный источник также играет роль различительного признака между аффрикатами */ц, ч/* и фрикативными */с, ш/*

(*/сок — цок, куча — куща/*) или разделителя фрикативных сегментов (*/сцена — сено/*). Глухие конечные согласные, как, например, в словах */суп, сук, сут/* (*сут* — разговорная реализация слова */суд/*), могут распознаваться только по спектрально-временным характеристикам речевого взрыва. Удаление взрыва конечного согласного обычно приводит к потере ощущения присутствия согласного. Аналогично, речевой взрыв маркирует присутствие глухого согласного перед гласным, например, при противопоставлении слов */почки — очки/*.

В исследованиях свойств речевых взрывов, порождаемых импульсным источником, исходят из представления о том, что в момент речевого взрыва его спектр наиболее точно отражает форму речевого тракта. В поисках инвариантов места артикуляции в [3–5] спектральные характеристики речевого взрыва изучались с использованием синтезированных и реальных звукосочетаний. В [6] было найдено, что амплитуда спектра взрыва в полосе частот выше 2500 Гц и в области частот второй и третьей форманты содержит информацию для определения места артикуляции. Отмечалось большое разнообразие спектральных характеристик речевого взрыва и их зависимость от последующего гласного.

Энергия речевого взрыва во много раз меньше энергии гласных и даже фрикативных, что создает определенные трудности в регистрации речевого взрыва и анализе его параметров объективными методами. Однако это не мешает субъективному восприятию речевого взрыва. В экспериментах с синтезированными слогами */da, ta/* было обнаружено, что если интервал времени ΔT_{vot} между началом речевого взрыва и началом последующего за ним гласного (так называемый Voice Onset Time) менее 20 мс, то в слуховом отделе коры головного мозга обезьяны наблюдается всплеск активности, соответствующий моменту речевого взрыва, а при длинных ΔT_{vot} (40–60 мс) к этому всплеску добавляется импульс, соответствующий началу гласного [7]. Электроэнцефалограммы и функциональная магнито-резонансная томография обнаруживают активность слуховой зоны коры головного мозга человека синхронно с импульсным взрывом, демонстрируя отчетливую разницу между слогами */ta/* и */da/* [8–10]. В [11] наблюдалось предпочтение правого уха при одновременном прослушивании слогов */pa/* и */ba/*, что указывает на важную роль левого полушария головного мозга в анализе речи на английском языке.

Величина ΔT_{vot} определяется координацией движения артикуляторов и голосовых складок. Эта координация зависит от многих факторов, в том числе и от состояния нервной системы. Так, было обнаружено, что по ΔT_{vot} можно диагностировать такие заболевания, как афазия, апраксия, дизартрия, болезнь Паркинсона, инсульт [12–14].

Присутствие голосового возбуждения на интервале смычки, который разделяет звонкие и глухие взрывные согласные */б, д, г/* и */п, т, к/*, например, в минимальных парах слов */почка — бочка, тыл — был, кора — гора/* может быть обнаружено по периодическим колебаниям в диапазоне частот основного тона 60–350 Гц. Однако по разным причинам этот признак неустойчив. При артикуляции глухих взрывных, в зависимости от координации артикуляторных движений и скорости разведения голосовых складок, в начале глухой смычки некоторое время могут наблюдаться затухающие колебания складок заметной длительности. На интервале звонкой смычки, наоборот, голосовой источник может прекратиться вследствие критического нарастания давления в ротовой полости. В телефонных каналах подавляется амплитуда именно в полосе частот основного тона. Поэтому необходимо иметь возможность детектирования признака звонкости/глухости другими средствами. Было обнаружено, что в некоторых языках в качестве такого признака можно использовать разницу распределения значений ΔT_{vot} для глухих и звонких взрывных согласных.

Одно из первых измерений ΔT_{vot} было выполнено на сонограммах речевого сигнала [15], где оценивались также интервалы времени между началом речевого взрыва и экстремумами осциллограммы вплоть до третьего импульса голосового возбуждения. Дальнейшие исследования обнаружили зависимость этого параметра от языка [16–18]. Обзор исследований ΔT_{vot} представлен в [19, 20]. Для некоторых языков среднее значение ΔT_{vot} у глухих взрывных в несколько раз больше среднего значения у звонких взрывных. В шведском языке средняя длительность ΔT_{vot} глухих взрывных составляет около 120 мс, а звонких — около 20 мс [21]. В английском языке средняя длительность ΔT_{vot} для глухих взрывных составляет около 70 мс, а звонких — около 15 мс [22]. Исследования ΔT_{vot} обычно выполняются на изолированных слогах, где этот эффект проявляется наиболее заметно. Но в английском языке характеристики ΔT_{vot} подтверждаются и для слитной речи [23].

Существуют языки, в которых практически отсутствует разница ΔT_{vot} между звонкими и взрывными. Так, в [15] сообщается, что, в отличие от английского языка, ΔT_{vot} начальных глухих взрывных испанского языка в несколько раз меньше (средние значения */p, t, k/* — 10, 15 и 25 мс в испанском языке против 58, 70 и 80 мс в английском языке). Такие малые значения ΔT_{vot} соответствуют звонким взрывным в английском языке, а в испанском языке ΔT_{vot} звонких взрывных приписываются отрицательные значения. При этом необходимо отметить, что часто используемое в литературе понятие отрицательного значения ΔT_{vot} противоречит определению ΔT_{vot} как интервала времени между речевым взрывом и первым импульсом голосового источника ***после*** взрыва.

В русском языке на материале бессмысленных звукосочетаний СГСГСГ (С — согласный, Г — гласный) были найдены численные значения ΔT_{vot} для различных артикуляторных событий [24]. Отмечалось, что среднее значение ΔT_{vot} согласных */n, m/* и */к/* в заударной позиции находится в пределах 14–17 мс, что существенно меньше, чем в английском языке. В предударной позиции ΔT_{vot} согласного */к/* заметно больше — в среднем около 43 мс, причем нередко появляются повторные речевые взрывы. Было также обнаружено, что на величину ΔT_{vot} в русском языке сильно влияет место артикуляции согласного — она возрастает от */n/* до */к/*. Величина ΔT_{vot} изменяется пропорционально длительности последующего гласного и увеличивается для мягких согласных (в два-три раза между */m/* и */m`/*). Установлена также зависимость ΔT_{vot} от длительности смычки, предударного или заударного положения согласного.

Эксперименты по восприятию речи в условиях маскирующего шума или низкочастотной фильтрации показали, что распознавание голосового возбуждения больше зависит от частоты основного тона, чем от ΔT_{vot} [25]. В большинстве экспериментов по оценке ΔT_{vot}, в которых установлена существенная разница между глухими и звонкими взрывными, речевой материал состоял из коротких изолированных бессмысленных звукосочетаний, тогда как установлена сильная зависимость ΔT_{vot} от числа слогов в слове [26]. На основе обзора результатов измерения ΔT_{vot} для 19 языков, в [19] сделан вывод о том, что ΔT_{vot} не может служить единственным параметром, определяющим различие между звонкими и глухими взрывными согласными. Анализ процессов речеобразования также приводит к выводу, что ΔT_{vot} зависит от языка, пола, объема легких, физиологического и психологического состояния человека, возраста, скорости речи, лингвистических условий (тип последующего гласного, число слогов во фразе), четкости артикуляции, громкости, частоты основного тона. Это заставляет усомниться в том, что ΔT_{vot} является инвариантным признаком, классифицирующим смычку как звонкую или глухую.

Автоматическое детектирование импульсного речевого взрыва наталкивается на серьёзные трудности, связанные с его кратковременностью и малой энергией. Такие его характеристики не только увеличивают погрешность оценки времени возникновения речевого взрыва, но и порождают множество ложных срабатываний. Поэтому прилагаются усилия по разработке помехоустойчивых методов автоматической его оценки [27–30]. Алгоритмы определения момента речевого взрыва действуют как во временной, так и в частотно-временной области, где применяются методы SVM, HMM, оператор Тигера-Кайзера, преобразование Гильберта или машинное обучение. При этом оказывается, что для детектирования импульсного речевого взрыва необходима предварительная сегментация на смычку и гласный с помощью автоматического распознавания речи, либо нужно использовать ограниченный тестовый материал, например, начальные взрывные изолированных слов. При таких условиях в [29–31] найдено, что в среднем около 60–90% оценок ΔT_{vot} обладают погрешностью менее 5 мс.

В восприятии речи импульсный источник выполняет несколько функций: сам факт присутствия речевого взрыва после смычки детектирует начальные и конечные взрывные, противопоставляя их гласным и фрикативным; различает глухие фрикативные и аффрикаты; содержит информацию о месте артикуляции; может указать на тип смычки — звонкую или глухую. Свойства импульсного источника относительно этих функций мало исследованы, тогда как они могут способствовать повышению эффективности решений в задачах автоматического распознавания речи, верификации диктора, синтеза речи по тексту или диагностики заболеваний. В данной работе характеристики импульсного источника в акустике речи изучаются на примерах артикуляторных событий, содержащих переход от смычки к гласному.

2. МАТЕМАТИЧЕСКИЕ МОДЕЛИ

Кратковременный сегмент речевого сигнала, который возникает в момент раскрытия смычки, порождается импульсным источником возбуждения акустических колебаний в речевом тракте. В терминологии речевых исследований этот сегмент называется речевым взрывом. Амплитуда импульсного источника определяется давлением в ротовой полости перед смычкой в момент раскрытия смычки и местом наибольшего сужения. Математические модели и измерения реальных аэродинамических процессов описываются в [32, 33]. На амплитуду импульсного источника также влияет средняя площадь голосовой щели, которая минимальна на сегменте звонкой смычки, но для глухих взрывных достигает величин, сопоставимых с минимальной площадью речевого тракта гласных звуков. После раскрытия звонкой смычки импульсный источник появляется на фоне акустических колебаний, вызванных действием квазипериодического голосового источника. После раскрытия глухой смычки колебания голосовых складок возобновляются через некоторое время, которое зависит от координации движений голосовых складок и движений артикуляторов.

Эти процессы в рамках модели аэродинамики речевого тракта [32] проиллюстрированы на рис. 1, на котором показана динамика минимальной площади речевого тракта S_{vt}, средней площади голосовой щели S_{vs}, давления под голосовой щелью

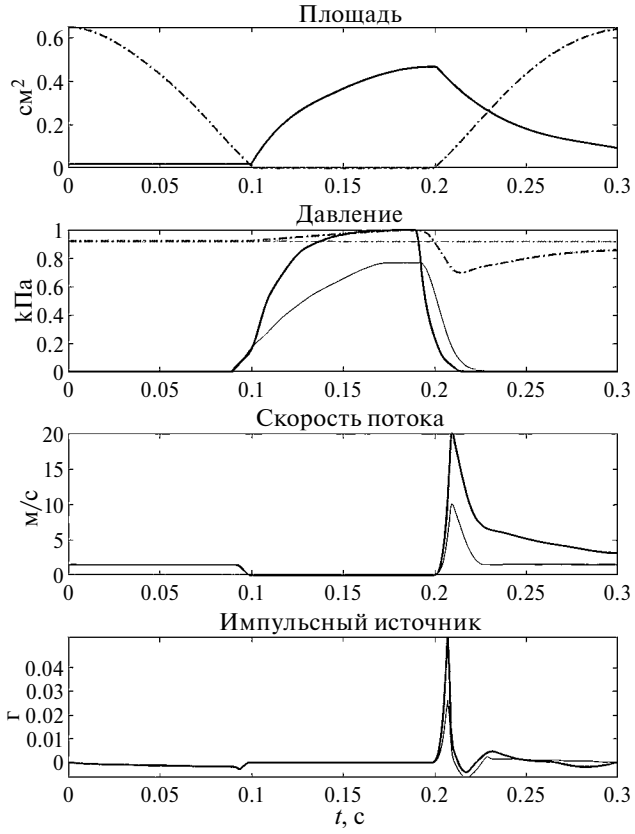

Рис. 1. Площадь S_{vt} (—), S_{vs} (—); давление P_{uvs} на глухой (-•-) и звонкой (-·-) смычке; давление P_{vt} на глухой (—) и звонкой (—) смычке; скорость V на глухой (—) и звонкой (—) смычке; импульсный источник W на глухой (—) и звонкой (—) смычке.

P_{uvs} и в ротовой полости P_{vt}, скорости воздушного потока V, а также представлен импульсный источник W в условиях глухой или звонкой смычки, расположенной между сегментами с голосовым возбуждением.

Несмотря на нарастание давления в ротовой полости, автоколебания голосовых складок во время звонкой смычки поддерживаются за счет податливости стенок речевого тракта и опускания гортани, что увеличивает объем ротовой полости. Но иногда в речевом сигнале наблюдается затухание низкочастотных периодических компонент к концу звонкой смычки. Это связано с таким накопленным давлением в ротовой полости, что условия существования автоколебаний голосовых складок нарушаются, и голосовой источник перестает действовать. Так порождается импульсный источник, соответствующий глухой смычке после сегмента звонкой смычки. Это может привести к ошибочной оценке звонкого взрывного согласного как глухого. Если скорость воздушного потока V на некотором интервале времени после раскрытия смычки достаточно велика для того, чтобы число Рейнольдса $\mathrm{Re} = V\rho_0 h / \mu$ превысило критическое значение, то создаются условия турбулизации воздушного потока, и в месте наибольшего сужения тракта возникает шумовой источник возбуждения. Здесь ρ_0 — плотность воздуха, μ — коэффициент вязкости, h — характерный геометрический параметр области сужения. Акустические характеристики импульсного источника на таком интервале аналогичны фрикативным звукам с соответствующим местом артикуляции. В зависимости от площади голосовой щели и ее координации с артикуляцией, длительность сегмента с шумовым возбуждением может быть короткой или длинной. В первом случае ΔT_{vot} принимает малые значения, а во втором вслед за импульсным источником располагается сегмент фрикативного шума, и такой источник в [17] было предложено называть аспиративным.

Спектральные характеристики импульсного источника на сегменте звонкой смычки включают характеристики импульса голосового возбуждения и фрикативных источников турбулентного потока воздуха на выходе голосовой щели и места наибольшего сужения в речевом тракте. Спектр производной по времени от воздушного потока после раскрытия смычки — широкополосный с падением к высоким частотам. Спектр фрикативного источника в речевом тракте также широкополосный, и содержит n гармоник с частотами $f_n = \mathrm{Sh}(\mathrm{Re})nV / d$, где Sh — число Струхаля ($\mathrm{Sh} \approx 0.2$), d — эквивалентный диаметр сужения. В экспериментах с артикуляторным синтезатором было обнаружено, что на восприятие фрикативных влияет не только первый резонанс турбулентного шума, но и второй, и, возможно, третий. Импульсный источник глухого речевого взрыва содержит также спектр турбулентного источника шума на выходе из голосовой щели. Максимум этого спектра мало зависит от формы речевого тракта, и находится на частоте около 700 Гц. Скорость воздушного потока после раскрытия смычки глухого или звонкого речевого взрыва заметно различается, как это видно на рис. 1. Поэтому и частота резонансов спектра фрикативного шума у звонкого взрыва f_n существенно ниже, чем у спектра глухого взрыва.

Место артикуляции глухих взрывных */n, m, к/* близко к месту артикуляции глухих фрикативных */ф, с, х/*, поэтому спектры глухого импульсного речевого взрыва (особенно, аспиративного) похожи на спектры глухих фрикативных. Аналогично, спектры взрыва на фоне звонкой смычки похожи на спектры звонких фрикативных. При этом спектр самих фрикативных зависит от места наибольшего сужения в речевом тракте, предшествующего и последующего гласного [34]. Одно из свойств волнового уравнения речевого тракта состоит в подавлении амплитуды колебаний на частоте резонанса, нуль собственной функции которого совпадает с положением импульсного источника, т.е. с местом наибольшего сужения в тракте. Благодаря этому

свойству спектры фрикативных */с, ш/* выглядят как отклики фильтров высокой частоты, а спектр */х/* — как отклик полосового фильтра. Аналогично, спектр импульсного речевого взрыва смычных согласных также искажается относительно спектра последующего гласного, на фоне которого происходит раскрытие смычки, и наиболее точно отражает место артикуляции взрывного согласного. Амплитуда некоторых пиков в спектре речевого взрыва (особенно, глухого или аспиративного) может быть подавлена вплоть до нуля или неразличимости от случайных флуктуаций. Если в окрестности резонанса гласного звука, последующего после речевого взрыва, в спектре взрыва не находится пик амплитуды, то это может быть следствием присутствия нуля, характерного для данного места артикуляции.

Акустические характеристики речевого тракта в момент глухого и звонкого речевого взрыва существенно различны. Если голосовые складки колеблются, то переменный поток воздуха через голосовую щель модулирует скорость потока в области смычки, изменяя форму импульсного источника. Спектр звонкого речевого взрыва зависит от фазы между колебаниями голосовых складок и динамикой площади речевого тракта в области смычки. Это дает основания для исследования возможности использования этого спектра для распознавания звонких согласных звуков. Относительно большая площадь голосовой щели во время глухой смычки приводит фактически к удвоению длины речевого тракта за счет трахеи и бронхов и, соответственно, к удвоению числа резонансов в заданном частотном диапазоне. При этом могут появиться низкочастотные резонансы, отсутствующие в передаточной функции тракта при автоколебаниях голосовых складок. Влияние подсвязочной области проявляется даже на гласных звуках в виде возникновения в спектре звуков пика на частотах около 400 и 1200 Гц, и падения амплитуды колебаний, когда резонансная частота речевого тракта пересекает 1200 Гц. Однако распределение резонансов такой акустической системы зависит еще и от граничных условий со стороны легких. Обычно легкие рассматриваются просто как некий объем, но многочисленные ветвления бронхиол создают сложную систему, импеданс которой, а следовательно, и граничные условия, могут обладать существенно разными свойствами в зависимости от геометрических размеров и параметров тканей легких [32]. Поэтому можно ожидать большого разнообразия характеристик глухого речевого взрыва в зависимости от индивидуальных особенностей диктора.

3. ЭКСПЕРИМЕНТЫ

Речевой взрыв смычки содержит информацию о разнообразных фонетических характеристиках речевого сигнала. В соответствии с этим нужно применять разные методы анализа, адекватные исследуемым характеристикам. Эти методы вытекают из математических моделей генерации речевого сигнала и, по возможности, должны учитывать известные свойства восприятия речи. Предполагается, что при субъективном анализе речи в слуховом аппарате человека используются амплитудный и фазовый спектр, а также детекторы амплитудных и частотных модуляций в различных масштабах времени.

В настоящей работе использовалась база данных, содержащая однозначные, двухзначные и трехзначные числительные, слитное произнесение 7-значного телефонного номера, а также некоторые служебные слова в задаче набора телефонного номера голосом. Запись велась одновременно на направленный микрофон и телефонную трубку с угольным микрофоном в обычной комнате. В записях принимали участие 24 мужчины и 15 женщин. Четыре человека, хорошо знакомые со свойствами речевого сигнала, выполняли разметку речевых сигналов на фонетические элементы. Далее обозначения фонетических элементов соответствуют системе из 127 типов артикуляторно-акустических элементов. В этой системе заглавные символы обозначают ударные гласные, прописные — безударные, а символ */ъ/* обозначает редуцированный твердый гласный.

Было обнаружено около 6320 появлений речевого взрыва Т! в сочетании с последующими фонетическими элементами */О, Ы, а, ы, ъ, В, Н, М/*, и около 4040 появлений речевого взрыва Д! в сочетании с последующими фонетическими элементами */А, И, Е, и, ъ, В/*. В сочетании с последующими фонетическими элементами */Е, И, и/* аспиративный взрыв T!h появлялся около 2020 раз, а аспиративный взрыв Дh около 2160 раз. При этом у женщин аспиративный взрыв как для глухой, так и для звонкой смычки, появлялся в несколько раз реже, чем импульсный взрыв. Аспиративный взрыв Пh был зарегистрирован более 4400 раз, тогда как взрыв Кh был найден около 100 раз.

В соответствии со свойствами внутреннего уха в данной работе с целью подъема амплитуд с ростом частоты выполнялось предыскажение речевого сигнала с коэффициентом 0.95. Затем сигнал разделялся гребенкой гамма-тон фильтров [35] на 128 каналов в шкале частот мел. Сигналы на выходе каждого фильтра подвергались преобразованию Гильберта с последующим вычислением фазового и логарифма амплитудного спектров. Некоторые характеристики речевого сигнала наиболее отчетливо проявляются в спектре групповой задержки, поскольку фазовые параметры меньше зависят от распределения энергии сигнала по частоте. Групповая задержка непрерывной фазы определяется как

$$\psi(\omega,t) = -2\pi\frac{\partial(\varphi(\omega,t))}{\partial\omega}, \qquad (1)$$

где $\varphi(\omega,t)$ — текущий спектр непрерывной фазы речевого сигнала, а $\psi(\omega,t)$ имеет размерность времени.

3.1. Оценка момента речевого взрыва

На первом этапе исследований использовались данные ручной разметки, что позволяет получить более или менее объективные сведения о свойствах речевых взрывов в различных условиях без погрешностей, связанных с автоматической разметкой. Затем были рассмотрены методы полуавтоматического определения характеристик речевого взрыва при условии, что известно начало смычки и начало гласного. Наконец, были выполнены оценки параметров речевого взрыва при автоматической разметке.

Малая энергия речевого взрыва по сравнению с шумом среды на сегменте предшествующей смычки и различие в длительности интервала между речевым взрывом и началом гласного затрудняют автоматический анализ. Это приводит к плохой помехоустойчивости энергетических критериев для детектирования речевого взрыва. Можно было бы использовать скорость нарастания энергии речевого взрыва, но простая производная речевого сигнала по времени также не помехоустойчива. Целесообразно применить так называемые динамические детекторы с различными масштабами времени, которые, возможно, являются аналогами детекторов амплитудных и частотных модуляций, обнаруженных в слуховой системе некоторых организмов [34, 36]. Анализ динамики звуков в слуховой системе человека выполняется в разных масштабах времени при сглаживании с постоянными времени от 1 до 300 мс. Считается, что кратковременный анализ на интервалах короче 40 мс выполняется в левом полушарии, а долговременный анализ — на интервалах 150–250 мс — в правом полушарии. Детекторы модуляций могут быть сформированы как во временной, так и в частотно-временной области. В последнем случае простейший детектор амплитудно-частотных модуляций представляется как

$$D(\omega,t,\tau_1,\tau_2,\Delta t) = \\ = \log[S(\omega_1,t,\tau_1) / S(\omega_2,t-\Delta t,\tau_2)], \quad (2)$$

где $S(\omega,t,\tau)$ — спектр речевого сигнала, сглаженный с постоянной времени τ, ω_1 и ω_2 — частотный диапазон анализа, τ_1,τ_2 — постоянные времени сглаживающих фильтров, а Δt — задержка или опережение. Детекторы такого типа обладают свойствами логарифмической производной и весьма устойчивы к шумам. Параметры $\tau_1,\tau_2,\Delta t$ должны соответствовать динамическим свойствам исследуемого параметра речевого сигнала. В частности, детектор речевого взрыва в полосе 1000–6000 Гц и постоянными времени сглаживания $\tau_1 = 1$ мс, $\tau_2 = 15$ мс и задержкой $\Delta t = -15$ мс обычно достаточно уверенно идентифицирует как звонкий, так и глухой речевой взрыв. Этот детектор может использоваться либо в виде суммарной энергии в заданной полосе частот 1000–6000 Гц

$$d_1(t) = \int_{\omega_{\min}}^{\omega_{\max}} D d\omega, \quad (3)$$

либо как двумерная функция $D_1(\omega,t) = D(\omega,t,\tau_1,\tau_2,\Delta t)$. В силу кратковременности, детектор $d_1(t)$ в состоянии также отслеживать импульсы голосового возбуждения на сегменте гласного, а огибающая детектора $d_2(t)$ с параметрами сглаживания $\tau_1 = 5$ мс, $\tau_2 = 25$ мс и задержкой $\Delta t = -40$ м реагирует на возрастание амплитуды гласного. На рис. 2 представ-

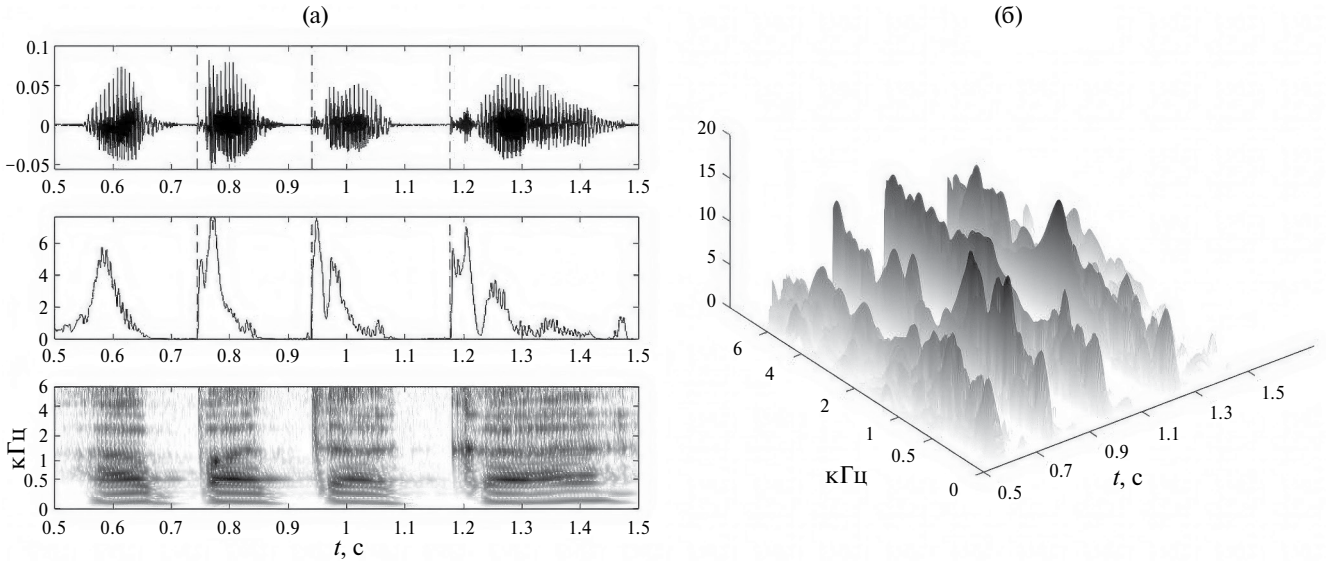

Рис. 2. Звукосочетание */апатакА/*. (а) — Сверху вниз: речевой сигнал, суммарный отклик кратковременного динамического детектора, сонограмма речевого сигнала; начало речевого взрыва — пунктир. (б) — Трехмерная сонограмма кратковременного динамического детектора.

лены осциллограммы звукосочетания */анатакА/*, моменты начала речевого взрыва, найденные путем субъективного анализа, сглаженная огибающая детектора d_1, и сонограмма речевого сигнала. Также показана трехмерная сонограмма кратковременного динамического детектора $D_1(\omega,t)$.

На этом рисунке видно, что и амплитудный спектр речевого сигнала, и спектр динамического детектора содержат пики, которые особенно четко выражены в спектре динамического детектора. Некоторые частоты этих пиков близки к резонансным частотам речевого тракта. В отличие от спектров звонкого речевого взрыва, и амплитудный спектр и спектр динамического источника глухого взрыва более похожи на спектры фрикативного */ф/* в сегменте */na/*, */c/* в сегменте */ma/* и */x/* в сегменте */кА/*. При этом особенно выраженный пик на частоте около 1500 Гц характерен для */x/*, произнесенного на фоне артикуляции с последующим переходом к гласному */a/* [34]. Эффективность детектора d_1 при условии, что известна окрестность истинного начала глухого или звонкого речевого взрыва, демонстрируется рис. 3 для разнообразных сочетаний взрывов Т! и Д! с последующими гласными для мужских голосов, которые представлены в исследуемой базе данных.

Пик кратковременного детектора сдвинут относительно начала взрыва примерно на 8 мс. Такое распределение оценок представляется вполне правдоподобным, учитывая, что между началом взрыва и наибольшей скоростью его изменения проходит некоторое время.

В силу кратковременности, динамические детекторы типа $D(\omega,t,\tau_1,\tau_2,\Delta t)$ реагируют на любые спектрально-временные модуляции, в частности, и на импульсы голосового источника на сегментах гласных звуков, как это видно на рис. 2. Это приводит к множеству ложных срабатываний детектора взрыва, которые необходимо фильтровать, исходя из оценок начала и конца смычки. Нас интересуют только те пики кратковременного детектора, которые располагаются перед началом гласного звука после смычки.

Прежде всего, необходимо выделить интервал времени, на котором, предположительно, присутствует речевой взрыв. Как упоминалось в Введении, с этой целью обычно используется сегментация речевого сигнала методами автоматического распознавания речи, где выполняется сегментация речевого сигнала на смычку и последующий сегмент — гласный, фрикативный, назальный, другая смычка или звуки */В, Л, Р/*. При оценке положения и параметров речевого взрыва в конструкции “смычка-гласный” целесообразно одновременное использование пиков кратковременного и долговременного детектора и моментов пересечения функциями d_1 и d_2 некоторого плавающего порога. Для каждой функции этот порог увеличивается, начиная с малого значения до такой величины, когда d_1 или d_2 его пересекают. Оба детектора реагируют не только на речевой взрыв, но и на любое возрастание энергии.

В отличие от рис. 3, на рис. 4 показаны распределения разности оценки начала речевого взрыва по полу-автоматической разметке для разнообразных сочетаний с гласными глухой и звонкой переднеязычной смычки, взрывов Д!, Т! и аспиративного взрыва Th, когда заданы начало смычки и начало гласного. Здесь объединены данные для мужских и женских голосов, а также для направленного микрофона и телефонной трубки. Общее число использованных сегментов составляло от 4600 до

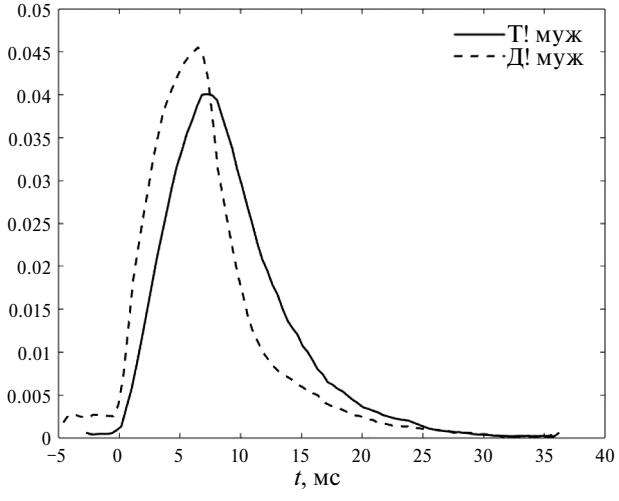

Рис. 3. Распределения разности оценок положения пика детектора d_1 и момента начала импульсного источника по данным ручной разметки.

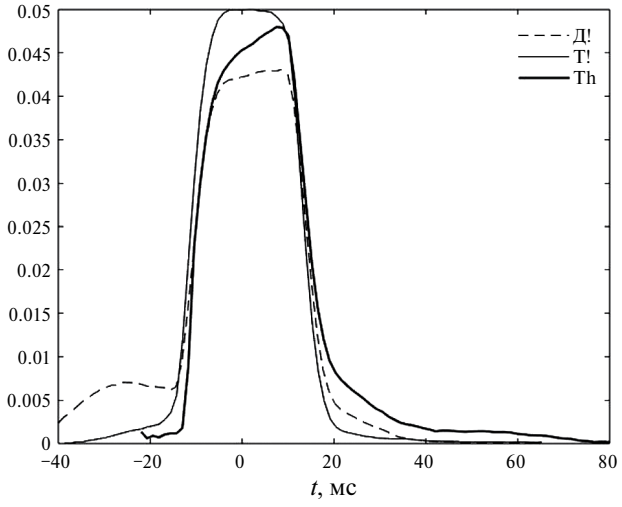

Рис. 4. Полуавтоматическая оценка начала импульсного речевого взрыва на сегментах */смычка-гласный/*.

11700. Показанные распределения примерно равномерны относительно истинного начала речевого взрыва в диапазоне ±10 мс, т.е. появились отрицательные оценки относительно начала взрыва.

Другой способ детектирования речевого взрыва использует сходство собственных функций спектров взрыва конкретного согласного с текущим спектром речевого сигнала. В отличие от неспецифических динамических детекторов, такие детекторы специфичны для каждого сочетания согласного и последующего гласного. В [34] фрикативные хорошо распознавались в пространствах собственных функций их спектров, найденных с помощью иерархического метода κ-средних (*k-means*). В этом алгоритме минимизируется сумма расстояний в евклидовой метрике между элементами кластера и его центроидом. В данной работе использовалось по 3 собственных функции для каждого сочетания речевого взрыва с последующими гласными.

В качестве меры сходства использовался модифицированный коэффициент Коши-Буняковского, чувствительность которого зависит от постоянной составляющей. Поэтому эту составляющую нужно вычитать как из собственных функций, так и из текущего спектра

$$K_{cb}^{(m)}(t) = \frac{\sum_{k=1}^{N} f_1(t,\omega_k) f_2(\omega_k)}{\sqrt{\sum_{k=1}^{N} (f_1(t,\omega_k) + c_r)^2 \sum_{k=1}^{N} f_2(\omega_k)^2}} \qquad (4)$$

где $f_1(t,\omega_k) = S(t,f_k) - \bar{S}(t,f_k)$, $f_2(\omega_k) = \psi_m(f_k) - \bar{\psi}_m(f_k)$, $c_r = \varepsilon k^2$, $\varepsilon = 10^{-4}$ — малая относительно максимума амплитуды $S(t,f_k)$ величина, ψ_m и $\bar{\psi}_m$ — m-я собственная функция спектра речевого взрыва и ее среднее значение, $S(t,f_k)$ и $\bar{S}(t,f_k)$ — текущий спектр и его среднее значение, $N = 128$ — число отсчетов спектра, $m = 1, 2, 3$ — номер собственной функции. Текущая мера сходства взрыва, например, Т!, определяется как максимум по всем m

$$K_{cb}^{(T!)}(t) = \max[K_{cb}^{(m)}(t)], m = 1,2,3. \qquad (5)$$

На рис. 5 показано распределение разности между положением пика функции $K_{cb}^{(Th)}(t)$ и ручной разметкой начала взрыва Th. Здесь пик наибольшего сходства смещен относительно начала взрыва, а само распределение заметно шире показанного на рис. 4. Поскольку в обоих способах оценок используются параметры разной физической природы, то при их совместном применении можно ожидать улучшения точности детектирования момента начала речевого взрыва.

На рис. 6 представлены отклики кратковременного (—) и долговременного (—) первичного детектора амплитудных модуляций, а также коэффициент максимального сходства текущего спектра речевого сигнала и множества собственных функций аспиративного взрыва Th в слове */cmo/*. Вертикальные линии отмечают возможное положение начала взрыва, найденного анализом первичных детекторов (—), и коэффициента максимального сходства (—). Согласованное решение о начале речевого взрыва показано на осциллограмме речевого сигнала.

Этот пример показателен во многих отношениях, поскольку он демонстрирует возможность отсева ложных решений путем сравнения оценок, полученных разными способами. Первичные детекторы неспецифичны к фонетическому содержанию сегментов речевого сигнала и реагируют на любые амплитудные модуляции, в том числе и на переход из одного артикуляторного состояния

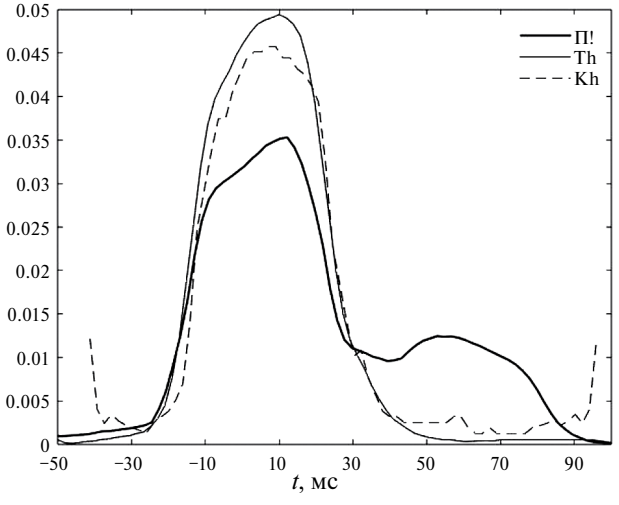

Рис. 5. Рассогласование между положением максимума сходства собственных функций П!, Th, Kh и ручной разметкой начала речевого взрыва.

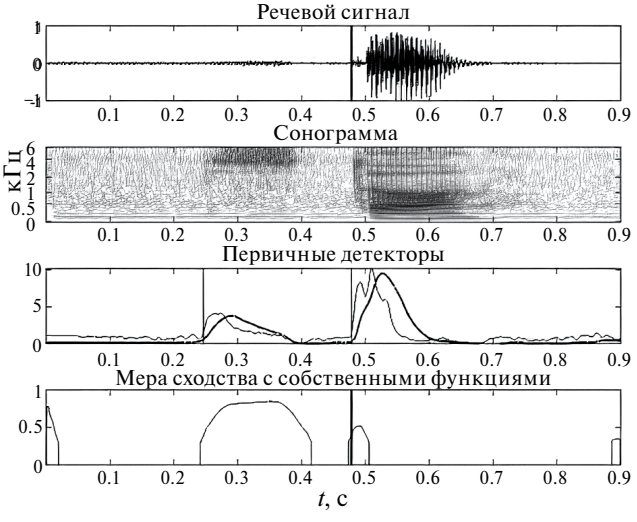

Рис. 6. Автоматическая оценка начала взрыва в слове */cmo/*.

в другое. Поэтому на рис. 6. начало фрикативного /*с*/ регистрируется первичными детекторами как потенциальное начало речевого взрыва. С другой стороны, коэффициент $K_{cb}^{(Th)}$ максимального сходства с собственными функциями взрыва Th высок, поскольку место артикуляции и спектральные свойства этого взрыва и фрикативного близки. Однако в этом случае не предлагается оценка о начале речевого взрыва, поскольку не удовлетворяются условия положения экстремума этого коэффициента относительно начала согласного. В итоге оказывается, что оценки начала речевого взрыва по первичным детекторам и коэффициенту практически совпадают, и совместное решение показано на панели речевого сигнала.

3.2. Различие между глухой и звонкой смычкой

Наряду с распознаванием периодических колебаний на интервале смычки, в речевых исследованиях большое внимание уделялось ΔT_{vot} как источнику информации о глухой или звонкой смычке, но статистически достоверных сведений об успешном использовании этого параметра в автоматическом распознавании русской речи найти не удалось.

В используемой базе имеется большое число появлений звонкой смычки согласного /*д*/, но количество представителей взрывов звонких взрывных /*б, г*/ слишком мало для сколько-нибудь достаточной статистики. Ориентировочные оценки ΔT_{vot} этих звуков можно получить, используя небольшую коллекцию звукосочетаний $\text{ГС}_1\text{ГС}_2\text{ГС}_3$, где в каждом звукосочетании мужским голосом произносилась одна и та же гласная из списка /*а, э, о, у, ы, и, е*/, а согласные произносились как /*п, т, к*/ или /*б, д, г*/. Например, произносились звукосочетания /*абадагА*/, /*упутукУ*/ и т.д. Средние оценки ΔT_{vot} этих согласных по всем гласным представлены в табл. 1. Эти данные как будто подтверждают ранние измерения на бессмысленных звукосочетаниях [24], и интервал между началом речевого взрыва и началом последующего гласного у глухих взрывных заметно больше, чем у звонких.

Таблица 1. Средние оценки ΔT_{vot}, мс.

Б	Д	Г	П	Т	К
13.9	19.4	25	25.9	41.1	51.1

Однако результаты экспериментов с реальными словами по оценке ΔT_{vot} взрывов глухих и звонких согласных /*т, д*/ по всем сочетаниям с последующими сегментами с голосовым возбуждением оказались существенно отличными от тех, которые были получены в [24], и от данных для других языков (рис. 7 и табл. 2). В частности, в [37] сообщается, что в английском языке ΔT_{vot} глухих взрывных

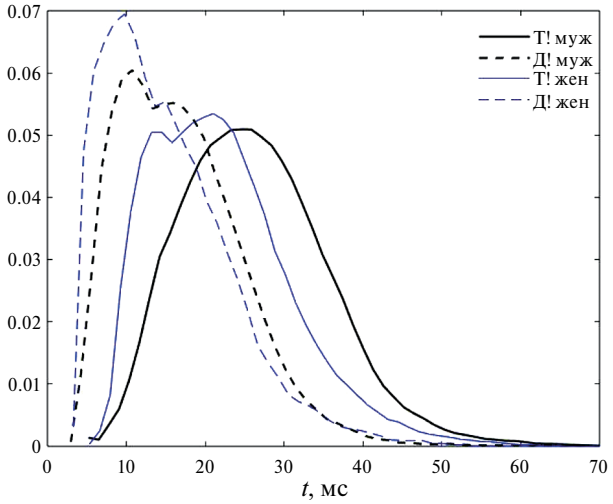

Рис. 7. Распределения ΔT_{vot} речевых взрывов глухих и звонких согласных; ручная разметка.

у женских голосов больше, чем у мужских голосов. Измерения в нашей базе данных демонстрируют обратное соотношение для всех взрывных согласных, а ΔT_{vot} глухих и звонких взрывов с заметным фрикативным сегментом у женских голосов отличаются мало.

В табл. 2 наиболее вероятные значения распределений ΔT_{vot} речевых взрывов Т! и Д! мужских голосов различаются примерно в полтора-два раза, но абсолютная разница составляет всего около 10 мс. У женских голосов ΔT_{vot} практически не отличаются ни для строго импульсных, ни для аспиративных взрывов. Как видно на рис. 7, распределения для речевого взрывов заметно перекрываются, и гипотеза о присутствии голосового источника на предыдущем сегменте оправдывается лишь в том случае, когда ΔT_{vot} импульсного источника меньше 17 мс для женских голосов и 25 мс для мужских голосов. Это накладывает жесткие ограничения на погрешность автоматического определения ΔT_{vot} в задачах речевых технологий. Обнаружено также, что распределения ΔT_{vot} аспиративных взрывов Th и Дh практически полностью перекрываются, что не позволяет судить о типе предшествующей смычки.

Таблица 2. Наиболее вероятное значение ΔT_{vot} импульсного и аспиративного взрывов мужских и женских голосов по результатам ручной разметки, мс.

Тип речевого взрыва	Мужчины	Женщины
Т!	21	14
Th	37	39
Д!	11	11
Дh	25	32

Согласно литературным данным, упомянутым в Введении, среднее значение ΔT_{vot} звонких взрывов в германских языках (15 мс) в несколько раз меньше, чем глухих взрывов (70 мс). По результатам измерений в использованной базе данных, среднее значение ΔT_{vot} глухих согласных */n, т, к/*, усредненное по мужским и женским голосам, типам микрофона и типам речевого взрыва, в разговорном русском языке оказывается существенно меньше, чем в германских языках (рис. 8). Максимум распределения ΔT_{vot} согласного */к/* приходится на 38 мс, что близко к среднему значению этого параметра для предударных и заударных позиций по [24].

На значение ΔT_{vot} влияют параметры артикуляторных органов и координация их движений. В этом отношении артикуляция согласных */к, г/* отличается от артикуляции губных и переднеязычных */п, б, т, д/*. В первом случае вовлечены относительно большие массы с заметной инерционностью, тогда как движения губ и кончика языка более быстрые, и координация их движений требует меньшего времени. На рис. 8 видно, что распределения ΔT_{vot} между */n, m/* различаются незначительно, а распределение для */к/* заметно отличается от распределения для */m/* и, особенно, для */n/*. Это может послужить признаком места артикуляции согласного */к/*. Для оценки возможности использования распределения ΔT_{vot} согласных */б, г/* при детектировании голосового источника на предшествующей смычке нужны дополнительные эксперименты со звонкими взрывными согласными.

Похоже, что в разговорном русском языке распределения ΔT_{vot} для глухих и звонких смычек в словах и фразах настолько пересекаются, что для детектирования присутствия голосового возбуждения на смычке этот параметр оказывается малоинформативным. Во Введении отмечалось аналогичное свойство испанского языка. Поэтому представляется целесообразным рассмотреть, помимо ΔT_{vot}, другие параметры импульсного взрыва с целью детектирования голосового источника на смычке.

Параметры речевого взрыва глухой и звонкой смычки различаются в области низких и высоких частот как в амплитудном спектре, так и в спектре групповой задержки. Величина $E_{\text{hl}} = E_{\text{high}} / E_{\text{low}}$ характеризует относительную энергию в области частот основного тона, связанную с присутствием голосового источника. Здесь E_{high} — средняя энергия в полосе частот 2500–6000 Гц, а E_{low} — средняя энергия в полосе 50–500 Гц на интервале времени $[T_{b\max}, T_{\text{vow}} - 2\text{ мс}]$, где $T_{b\max}$ — момент максимального значения импульсного источника, а T_{vow} — момент начала гласного. Распределения этого отношения для противопоставления звонкого Д! и глухого взрыва Т! между согласными */m, д/* для мужчин и женщин по данным ручной разметки базы числительных русского языка представлены на рис. 9. Было также обнаружено, что распределение отношения энергии E_{hl} мало зависит от типа взрыва — сосредоточенного (Т!, Д!) или с заметной долей фрикативного сегмента (Th, Dh).

В речевых сигналах, записанных через телефонную трубку, распределения E_{hl} в амплитудно-частотном спектре для глухих и звонких взрывов полностью перекрываются, что связано с подавлением частот ниже 300 Гц, принятое в проводных телефонных системах. Спектр групповой задержки $\psi(\omega,t)$ меньше зависит от свойств амплитудно-частотного спектра, и относительная энергия в области высоких и низких частот в сигналах от телефонного микрофона лучше детектирует

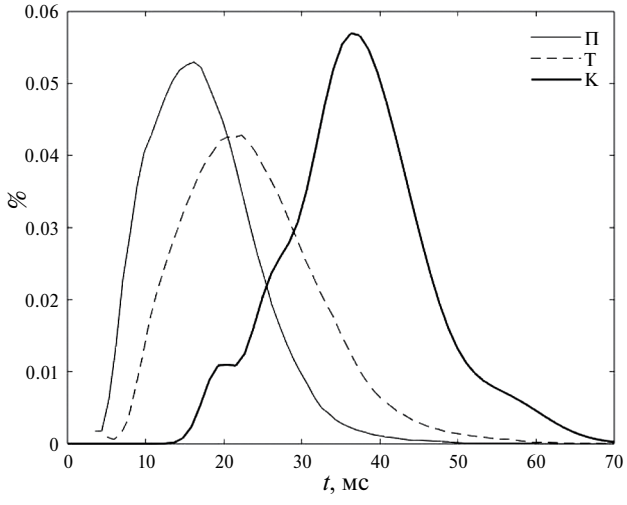

Рис. 8. Распределение ΔT_{vot} глухих взрывных согласных по результатам ручной разметки.

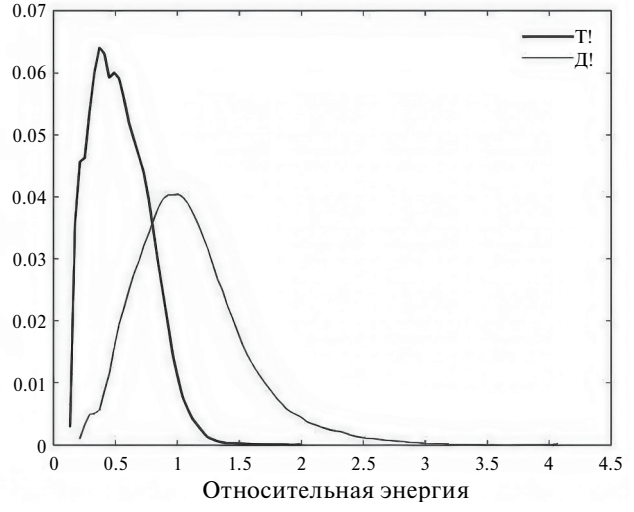

Рис. 9. Распределения относительной энергии импульсных взрывов Т! и Д! в амплитудном спектре; направленный микрофон.

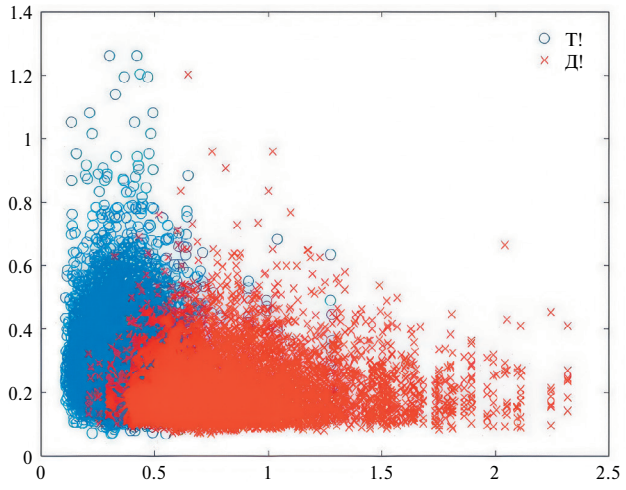

Рис. 10. Амплитудные и фазовые отношения на глухих и звонких речевых взрывах. По оси абсцисс — относительная энергия амплитудного спектра, по оси ординат — относительная энергия спектра групповой задержки.

присутствие голосового источника. С этой целью в каждый момент времени $\psi(\omega,t)$ нормируется путем вычитания низкочастотной компоненты, т.е. спектра, сглаженного по 17 отсчетам гамма-тон фильтров. Затем находится среднее значение абсолютной величины групповой задержки ψ_{low} в диапазоне частот 50–500 Гц и ψ_{high} в диапазоне частот 500–4000 Гц и вычисляются распределения отношений $\psi_{hl} = \psi_{high} / \psi_{low}$ в окрестности взрыва смычек согласных. При этом оказалось, что распределения ψ_{hl} мужских и женских голосов практически не отличаются. Совместное распределение относительной энергии в амплитудном спектре и спектре групповой задержки на сегменте речевого взрыва для мужских и женских голосов лучше обеспечивает разделимость признаков звонкости/глухости, чем одномерные статистики (рис. 10).

3.3. Место артикуляции

Импульсный источник возникает в момент времени, когда спектральные характеристики речевого тракта наиболее точно соответствуют месту артикуляции. Но амплитуда колебаний на резонансных частотах тракта в момент речевого взрыва зависит также и от расположения нулей соответствующих собственных функций резонансов и амплитуды максимума спектра турбулентного шума на этих частотах. Поэтому пики амплитудного и фазового спектра речевого взрыва, а также спектра кратковременного детектора, не обязательно соответствуют резонансам тракта и заметно различаются у глухого и звонкого речевого взрыва. Разнообразие этих спектров представлено на рис. 11 на сегментах звукосочетаний */абадагА/* и */апатакА/*, где показаны средний спектр кратковременного динамического детектора d_1 в окрестности пика [–5, 2.5] мс и средний спектр взрыва на интервале [0, 10] мс относительно начала речевого взрыва.

Простейший способ определения резонансных частот речевого тракта состоит в их идентификации по частотам локальных пиков амплитудного спектра. Это не очень точный метод, поскольку результаты зависят от многих факторов, таких как тип микрофона, помехи, реверберация помещения, но его удобно использовать для предварительного качественного анализа возможности детектирования места артикуляции по характеристикам взрыва. В табл. 3 представлены пики спектра звонкого речевого взрыва и спектра кратковременного детектора, а также пики спектра на удалении в 30 мс от начала гласного. Отсчеты этих спектров выполнялись по результатам ручной разметки. Как видно, наблюдается примерное соответствие пиков спектров взрыва и кратковременного детектора резонансам гласных от второго и выше. Спектры глухого взрыва заметно хуже идентифицируют резонансы речевого тракта, и в них присутствуют дополнительные пики.

Спектральные характеристики глухого речевого взрыва до некоторой степени соответствуют характеристикам глухих фрикативных, но распознавание глухих взрывов не столь эффективно. Глухие взрывные */n, m, к/* различаются в разных пространствах измерений параметров речевого взрыва. Так, */n/* заметно отличается от */к/* в распределениях частоты максимального пика спектра речевого взрыва в полосе 3–3.5 кГц и от */m/* в полосе 3–4 кГц (рис. 12). Взрыв согласного */к/* также разделяется от взрывов */n, m/* в распределении отношения средней энергии в полосе 0.8–3 кГц к энергии в полосе 3–6 кГц (рис. 13). Распределения на рис. 12 и 13 получены суммированием данных для всех видов взрыва, мужчин и женщин, направленного и телефонного микрофона.

Как было установлено в исследованиях фрикативных [34], при сравнении спектров в полном диапазоне частот 0–6 кГц необходимо учитывать их разнообразие, связанное с особенностями дикторов, влиянием условий регистрации речевого сигнала и окружающих звуков. С этой целью используется множество собственных функций спектров речевых взрывов каждого согласного (см. раздел 3.1). В отличие от фрикативных согласных, взрывные согласные характеризуются не только спектром взрыва, но и спектром отклика кратковременного динамического детектора. Оптимальное число собственных функций для каждого согласного в данной работе оказалось равным 3. В распределениях максимальной меры сходства $K_{cb}(t)$ собственных функций согласных */n, m, к/* и спектров их собственных взрывов и взрывов других согласных лучше всего распознается */n/* (рис. 14).

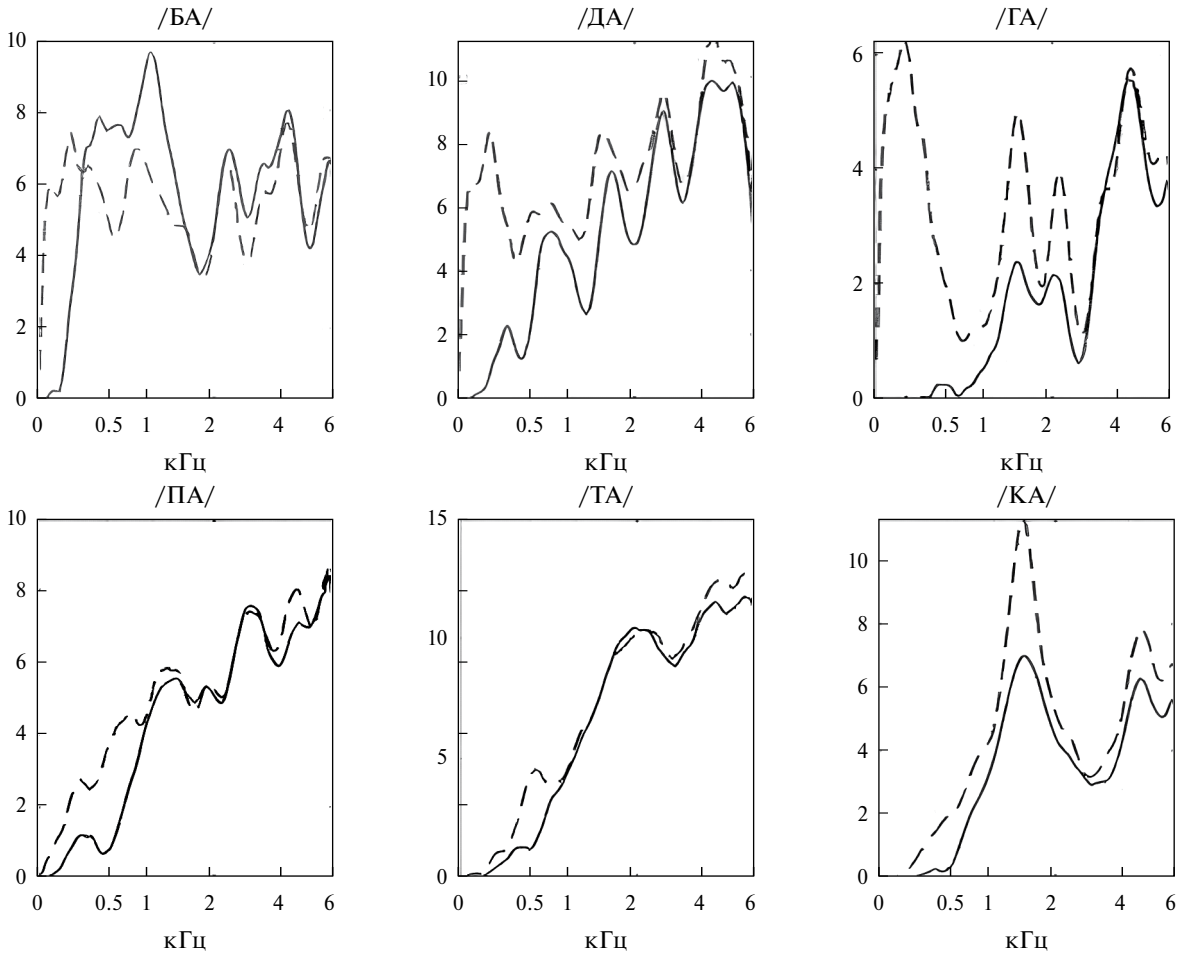

Рис. 11. Спектры звонких и глухих речевых взрывов: кратковременный динамический детектор (—); взрыв (---).

Таблица 3. Частота пиков спектра звонкого речевого взрыва, кратковременного детектора и гласного, кГц.

Форманты взрыва	Форманты детектора		Форманты гласного
0.87 1.45 2.38 3.36	0.96 2.5 3.2 4	БА	0.6 1.17 2.44 3.36
0.87 1.27 1.76 2.38	1.48 1.81 2.38 3.22	БЭ	0.47 1.48 2.33 3.09
0.79 1.45 1.85 2.38	1.48 1.81 2.38 3.29	БО	0.5 1.37 3.03
0.58 2.18 2.96 3.88	0.77 1.34 2.44 3.65	БУ	0.34 0.62 0.93 2.23 3.88
1.24 2.08 2.96 4.05	1.07 1.94 3.02 3.96	БЫ	0.29 1.45 2.13 3.29 3.96
1.48 2.49 3.29	0.85 1.94 2.55 3.22	БИ	0.3 0.82 1.52 2.18 3.16
0.79 1.52 2.78	0.72 1.6 2.72 4.13	ДА	0.47 1.52 2.66
0.56 0.85 1.2 2.72	0.87 2.9 4.21	ДЭ	0.35 0.93 1.72 2.66 3.88
0.79 1.37 3.58	1.41 2.28 3.43	ДО	0.37 0.9 1.37 2.55 3.29
0.96 1.41 2.44 3.36	0.9 1.45 2.44 3.36	ДУ	0.29 0.49 0. 9 1.34 2.44
0.93 1.6 2.28 3.43	1.81 3.51 4.65	ДЫ	0.32 0.93 1.6 2.49 3.43
1.37 2.13 2.96 3.65	0.79 2.13 2.84 3.58	ДИ	0.3 1.52 2.18 2.9 3.58
0.96 1.48 2.28	1.41 2.18 3.29 4.12	ГА	0.45 0.99 1.45 2.28 3.29
0.79 1.64	1.68 2.28 3.43 4.12	ГЭ	0.47 1.34 1.76 2.44 3.43
0.79 2.44 3.16	0.74 0.93 4.21	ГО	0.39 0.85 2.38 3.29
0.69 2.13 3.88	0.74 1.3 2.08 3.72	ГУ	0.39 0.85 2.38 3.29
1.52 2.13 3.36	1.56 2.38 4.21	ГЫ	0.31 1.56 2.08 3.29
0.87 1.17 1.81 3.43	1.14 3.03 3.43	ГИ	0.29 0.47 0.72 2.23 3.58

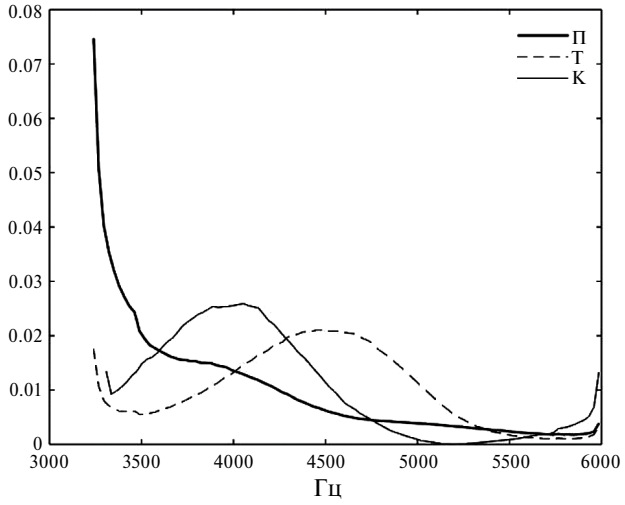

Рис. 12. Распределение частоты максимального пика спектра речевого взрыва для глухих взрывных.

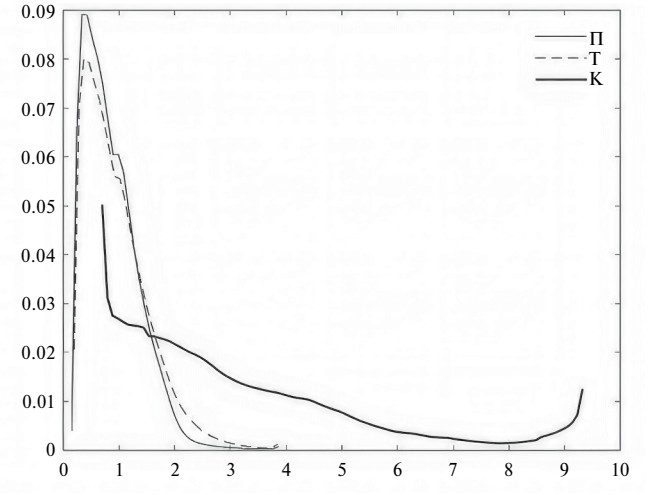

Рис. 13. Распределение отношения средней энергии в полосе 0.8–3 кГц к энергии в полосе 3–6 кГц.

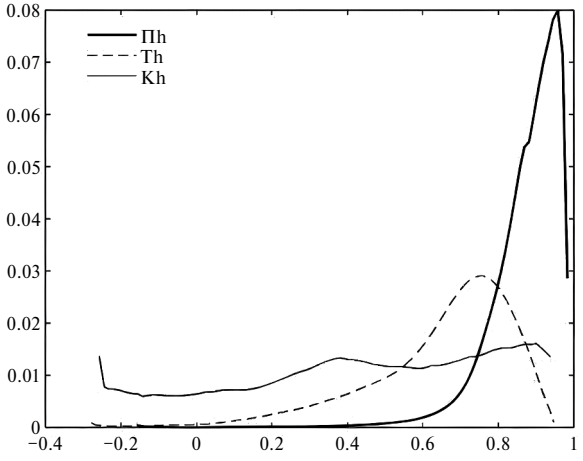

Рис. 14. Мера сходства собственных функций спектра речевого взрыва Пh (—) и спектров речевых взрывов Th и Kh.

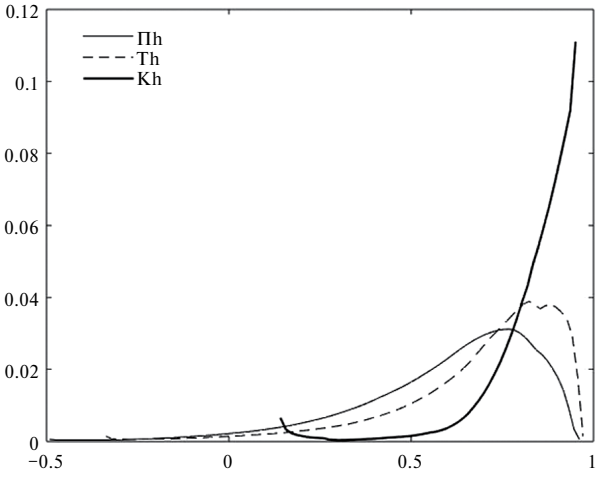

Рис. 15. Мера сходства собственных функций кратковременного детектора спектра речевого взрыва Kh (—) и спектров кратковременного детектора речевых взрывов Пh и Th.

Согласный /*к*/ лучше распознается в пространстве собственных функций спектра отклика динамического детектора (рис. 15).

4. ОБСУЖДЕНИЕ

При обсуждении результатов экспериментов необходимо выявить источники и размер погрешности, чтобы определить, насколько значимы сходство или различие наблюдаемых явлений. Оценка фонетических параметров речевого взрыва зависит от погрешности детектирования его начала. При этом в ряде случаев наблюдаются отрицательные оценки относительно ручной разметки, как показано на рис. 3, 4 и 5. Такие оценки могут возникнуть как в силу погрешности детектора, так и вследствие субъективного мнения разметчика речевого сигнала относительно начала речевого взрыва. Даже располагая исходной информацией в виде осциллограммы, сонограммы и поведения динамического детектора, в ряде случаев невозможно определить “истинное” начало речевого взрыва с точностью до 1 мс. В формировании нашей базы данных принимали участие четыре разметчика, что позволяет до некоторой степени снизить такую субъективность. По условиям разметки за начало импульсного источника принимается момент времени, в который амплитуда речевого сигнала на осциллограмме или сонограмме “заметно” превышает предшествующий фон. Один и тот же речевой взрыв разные разметчики могут пометить и как сосредоточенный, и как аспиративный. При этом разброс оценок разными разметчиками одного и того же события в среднем составлял около 5 мс. Это приводит к размытию распределений исследуемых параметров и увеличению области их

пересечения. В частности, распределения глухого и звонкого речевого взрыва на рис. 3 с учетом погрешности оказываются практически неразличимыми. Слишком большие положительные погрешности также являются следствием недостатков использованных методов анализа. Эти погрешности необходимо учитывать при автоматическом анализе определения вероятности “истинного” положения начала речевого взрыва.

Прежде, чем использовать акустические параметры речевого взрыва для оценки фонетических параметров речевого сегмента, необходимо детектировать сам взрыв, определив его начало и конец. При этом необходимо избавиться от ложных решений о присутствии речевого взрыва, что особенно сложно в силу его малой энергии относительно шумов акустической среды и помех в электронных каналах. Речевой взрыв смычки есть элемент артикуляторных событий, поскольку он возникает в процессе перехода от состояния смычки к последующему сегменту — гласному, фрикативному, назальному или другой смычке. Поэтому задача детектирования речевого взрыва является элементом более общей задачи детектирования последовательности артикуляторных состояний /(*смычка, пауза*) — *взрыв* — (*гласный, фрикативный, назальный, смычка, /В, Л, Р/*). Это означает предварительную сегментацию речевого сигнала на такие элементы. Если практическая задача состоит, например, в оценке координации артикуляторных движений с целью диагностики заболеваний нервной системы, то такая сегментация может быть выполнена средствами автоматического распознавания речи. Если же анализ параметров речевого взрыва выполняется для оценки места артикуляции согласного в интересах распознавания речи, то необходима сегментация речевого сигнала на базовые элементы */пауза, смычка, гласный, фрикативный, назальный/*. Есть основания полагать, что такая сегментация является элементом общей задачи распознавания речи.

Эксперименты по восприятию взрывных согласных показали, что, хотя роль речевого взрыва в распознавании места артикуляции ниже вклада формантных частот, но все же довольно заметна [1, 2]. В ранних работах по анализу спектральных характеристик речевого взрыва безуспешно пытались найти инварианты места артикуляции, поскольку в момент взрыва резонансные частоты речевого тракта должны наиболее точно отражаться в спектре речевого взрыва. Вскоре после изобретения сонографа было обнаружено, что треки формант перед смычкой и сразу после нее определяются взаимодействием согласного и окружающих его гласных. При этом разность частот второй и третьей форманты, например, в начале гласного и их значениями на стационарном сегменте гласного, зависит от места артикуляции согласного. На этой основе была сформулирована так называемая теория локусов, декларирующая, что линейное продолжение отсчетов этих формант на некоторое время в сторону смычки позволяет вычислить “асимптотические” значения формантных частот, однозначно определяющие место артикуляции [39–41]. Это явление определяется тем, что смычка и ее взрыв происходят на фоне формы речевого тракта, соответствующей последующему гласному. В [32] было показано, что при близкой к нулю минимальной площади поперечного сечения речевого тракта скорость изменения резонансных частот тракта значительно выше скорости изменения минимальной площади. Поэтому сопоставление формантных частот в момент речевого взрыва и формантных частот на стационарном участке может более отчетливо опознать взрывной согласный, чем при использовании формантных частот в начале гласного. Вместе с тем, анализ данных, представленных в табл. 3, показывает, что пики спектра речевого взрыва, спектра отклика кратковременного детектора и спектра гласного находятся в более сложных взаимоотношениях, чем простая разность.

Все известные методы оценки формантных частот характеризуются заметной погрешностью. Другой способ распознавания места артикуляции взрывного согласного состоит в использовании параметров спектра речевого взрыва с учетом известных свойств процессов речеобразования. Результаты представленных в данной работе исследований свидетельствуют о том, что детектирование голосового источника на смычке и определение места артикуляции речевого взрыва должно выполняться в совокупности пространств измерений. Так, информация о звонкости/глухости смычки содержится в относительной энергии между высокими и низкими частотами в амплитудных и фазовых спектрах речевого взрыва, а также в интервале времени ΔT_{vot} между началом взрыва и началом гласного. Распознавание места артикуляции взрыва может быть основано на сопоставлении потоков оценок меры сходства собственных функций спектров речевого взрыва и динамического детектора каждого взрывного согласного с текущими спектрами речевого сигнала. Вероятностное распределение ΔT_{vot} содержит информацию о месте артикуляции взрывного */к/*. Взрыв всегда происходит на фоне артикуляции последующего звука или нейтрального состояния в конце высказывания. Поэтому при оценке места артикуляции речевого взрыва нужно учитывать распределение пиков его спектра и спектра последующего звука, как это видно из табл. 3, и частоту максимального пика амплитудного спектра речевого взрыва в высокочастотной области.

Результаты экспериментов по оценке ΔT_{vot} в русском языке, описанные в данной работе, до некоторой степени противоречат данным [24],

полученным на материале бессмысленных звукосочетаний. Но оценки параметров речевого сигнала в бессмысленных звукосочетаниях могут оказаться неприменимыми к словам и слитной речи. Известно, что непрерывный поток осмысленной речи управляется другими критериями, чем артикуляция бессмысленных звукосочетаний. В зависимости от темпа речи, партитура электромиографических команд, подаваемых на лицевые мышцы и внутри-ротовые мышцы, может существенно меняться. Это связано с действием разнообразных критериев оптимальности, действующих в системе управления артикуляцией. Кроме того, возникают специфические эффекты коартикуляции, включающие ассимиляцию признака звонкости/глухости. Например в слитном словосочетании числительного “22” возможны варианты произнесения */два**д**цат′два/* и */два**т**цадва/*.

Обнаружено, что распределения ΔT_{vot} для звонких и глухих переднеязычных */т, д/* в русском языке сильно перекрываются, и между средними значениями существует лишь малая разница порядка 5 мс. Такое малое различие в распределениях ΔT_{vot} глухих и звонких взрывных в задачах распознавания речи требует высокой разрешающей способности алгоритмов анализа речевого сигнала, и вызывает сомнение в эффективности ΔT_{vot} в процессах субъективного восприятия звонкости/глухости смычки в русском языке. По некоторым данным, минимально воспринимаемая человеком разница в длительности составляет около 10 мс [41, 42]. В электрической активности коры головного мозга начало гласного звука не регистрируется, если ΔT_{vot} меньше 20-40 мс. Это означает, что ΔT_{vot} звонких смычек в русском языке, скорее всего, не детектируется, и в русском языке, как и в испанском, ΔT_{vot} служит по крайней мере слабым признаком звонкости/глухости смычки. Известно, что при обучении испанскому языку носителю английского языка приходится менять систему признаков для обозначения звонкости/глухости [43].

В исследованиях объективных различий между глухим и звонким речевым взрывом при обучении по размеченным сигналам следует учитывать исправляющую способность речевого кода на лексическом и семантическом уровне. Если в каком-то слове признак звонкости/глухости не служит смыслоразличительным элементом, то в слитной речи он может употребляться либо случайно, либо зависеть от каких-то других условий, например, действия критерия экономии энергии. Замена звонкого */д/* на глухой */т/* при произнесении слова */подол/* не приведет к ошибке восприятия, поскольку в словаре русского языка не существует слова */потол/*. Аналогично, замена */п/* на */б/* не приводит к смысловой ошибке. Ошибка в признаке звонкости начального взрывного в паре */бочка/* — */точка/* может быть скорректирована по смыслу фразы. Такая ситуация порождает эффект несоответствия буквенной записи речи и реального содержания акустических элементов и приводит к ошибкам в интерпретации результатов экспериментов при обучении системы автоматического распознавания речи.

5. ЗАКЛЮЧЕНИЕ

Положение речевого взрыва смычки в переходном процессе от смычки к гласному определяется в пространстве параметров кратковременного и долговременного детектора амплитудных модуляций и меры сходства с собственными функциями спектра. Интервал времени между началом речевого взрыва в русском языке является слабым признаком присутствия голосового возбуждения на смычке, но указывает на место артикуляции заднеязычных взрывных согласных. Информация о месте артикуляции взрывных согласных и присутствии голосового источника находится в пространстве таких параметров речевого взрыва, как спектр групповой задержки, частота максимального пика амплитудного спектра взрыва в высокочастотной области, отношение средней энергии спектра в области высоких и средних частот, а также в пространстве собственных функций спектра взрыва.

СПИСОК ЛИТЕРАТУРЫ

1. *Jongman A., Miller J.D.* Method for the location of burst-onset spectra in the auditory-perceptual space: A study of place of articulation in voiceless stop consonants // J. Acoust. Soc. Am. 1991. V. 89. N 2. P. 867–873.
2. *Bonneau A., Djezzar L., Laprie Y.* Perception of the place of articulation of French stop bursts // J. Acoust. Soc. Am. 1996. V. 100. P. 555–564.
3. *Dorman M.F., Studdert-Kennedy M., Raphael L.* Stop-consonant recognition: Release bursts and formant transitions as functionally equivalent, context-dependent cues // Perception & Psycophysics. 1977. V. 2. N 2. P. 109–122.
4. *Stevens K., Blumstein S.* Invariant cues for place of articulation in stop consonants // J. Acoust. Soc. Am. 1978. V. 64. P. 1358–1368.
5. *Blumstein S., Stevens K.* Acoustic invariance in speech production: evidence from measurements of the spectral characteristics of stop consonant // *J. Acoust. Soc. Am.* 1979. V. 66. P. 1001–017.
6. *Ohde R.N., Stevens K.N.* Effect of burst amplitude on the perception of stop consonant place of articulation // J. Acoust. Soc. Am. 1983. V. 74. P. 706–714.
7. *Steinschneider M., Fishman Y.I., Arezzo J.C.* Representation of the voice onset time (VOT) speech parameter in population responses within primary auditory cortex of the awake monkey // J. Acoust. Soc. Am. 2003. V. 114. N 1. P. 307–321.
8. *Sharma A., Dorman M.F.* Cortical auditory evoked potential correlates of categorical perception of voice-on-

set time // J. Acoust. Soc. Am. 1999. V. 106. N 2. P. 1078–1083.

9. *Steinschneider M., Volkov I.O., Noh M.D., Garell P.C., Howard III M.A.* Temporal encoding of the voice onset time phonetic parameter by field potentials recorded directly from human auditory cortex// J. Neurophysiol. 1999. V. 82. P. 2346–2357.
10. *Blumstein S.E., Myers E.B., Rissman J.* The Perception of Voice Onset Time: An fMRI Investigation of Phonetic Category Structure // J. Cognitive Neuroscience. 2005. V. 17. N 9. P. 1353–1366.
11. *Rimol L.M., Eichele T., Hugdahl K.* The effect of voice-onset time on dichotic listening with consonant-vowel syllables // Neuropsychologia. 2006. V. 44. N 2. P. 191–196.
12. *Auzou P., Ozsancak C., Hard R., Morris J., Jan M., Eueustache F., Hannequin D.* Voice onset time in aphasia, apraxia of speech and dysarthria: a review // Clinical Linguistics & Phonetics. 2000. V. 14. N 2. P. 131–150.
13. *Min S.N., Park S.J., Im J.N., Subramaniyam M.A.* Bayesian Model for Prediction of Stroke with Voice Onset // 3rd Int. Conf. on Advances in Mechanical Engineering (ICAME2020). IOP Conf. Series: Materials Science and Engineering. 2020. 912(6). 062003. https://doi.org/10.1088/1757–899X/912/6/062003
14. *Johansson I.-L., Samuelsson Ch., Müller N.* Consonant articulation acoustics and intelligibility in Swedish speakers with Parkinson's disease: a pilot study // Clinical Linguistics & Phonetics. 2023. V. 37. N 9. P. 845–865.
15. *Lisker L., Abramson A.* A cross-language study of voicing in initial stops: Acoustical measurements // Word. 1964. V. 20. P. 384–422.
16. *Ladefoged P., Madison I.* The sounds of the world's languages. Blackwell Publishing, 1996.
17. *Cho T., Ladefoged P.* Variation and Universals in VOT: Evidence from 18 Languages // J. Phonetics. 1999. V. 27. P. 207–229.
18. *Chodroff E., Golden A., Wilson C.* Covariation of stop voice onset time across languages: Evidence for a universal constraint on phonetic realization // J. Acoust. Soc. Am. 2019. V. 145 (1). EL109-EL115.
19. *Cho T., Whalen D.H., Docherty G.* Voice onset time and beyond: Exploring laryngeal contrast in 19 languages // J. Phonetics. 2019. V. 72. P. 52–65.
20. *Winn M.B.* Manipulation of voice onset time in speech stimuli: A tutorial and flexible Praat script // J. Acoust. Soc. Am. 2020. V. 147. N 2. P. 852–866.
21. *Fant G.* Speech and Sounds. MIT Press, 1973.
22. *Robb M., Gilbert H., Lerman J.* Influence of Gender and Environmental Setting on Voice Onset Time // Folia Phoniatrica et Logopaedica. 2005. V. 57. P. 125–133.
23. *Niyogi P., Ramesh P.* The voicing feature for stop consonants: Recognition experiments with continuously spoken alphabets. // Speech Commun. 2003. V. 41. P. 349–367.
24. *Сорокин В.Н.* Синтез речи. М.: Наука, 1992.
25. *Winn M.B., Chatterjee M., Idsardi W.J.* The roles of voice onset time and F0 in stop consonant voicing perception: Effects of masking noise and low-pass filtering // J. Speech Lang. Hear. Res. 2013. V. 56. N 4. P. 1097–1107.
26. *Yu V., De Nil L., Pang E.* Effects of age, sex and syllable number on voice onset time: Evidence from children's voiceless aspirated stops // Language and Speech. 2015. V. 58. N 2. P. 152–167.
27. *Stouten V., Van Hamme H.* Automatic Voice Onset Time Estimation from Reassignment Spectra // Speech Communication. 2009. V. 51. N12. P. 1194–1205.
28. *Hansen J.H.L., Gray Sh.S., Kim W.* Automatic voice onset time detection for unvoiced stops (/p/, /t/, /k/) with application to accent classification // Speech Communication. 2010. V. 52. P. 777–789.
29. *Lin Ch.-Y., Wang H.-Ch.* Automatic estimation of voice onset time for word-initial stops by applying random forest to onset detection // J. Acoust. Soc. Am. 2011. V. 130. N 1. P. 514–525.
30. *Sonderegger M., Keshe J.* Automatic Discriminative Measurement of Voice Onset Time // J. Acoust. Soc. Am. 2012. V. 132. N 6. P. 3965–3979.
31. *Prathosha P., Ramakrishnan A.G., Ananthapadmanabha T.V.* Estimation of voice-onset time in continuous speech using temporal measures // J. Acoust. Soc. Am. 2014. V. 136. N 2. EL122.
32. *Сорокин В.Н.* Теория речеобразования. 1985.
33. *Stevens K.N.* Acoustic phonetics. MIT, 1998.
34. *Сорокин В.Н.* Детекторы артикуляторных событий // Акуст. журн. 2020. Т. 66. № 1. С. 71–85.
35. *Patterson R.D., Holdsworth J.* A functional model of neural activity patterns and auditory images // Advances in Speech, Hearing and Language Processing. 1996. V. 3. P. 547–563.
36. *Сорокин В.Н., Чепелев Д.Н.* Первичный анализ речевых сигналов // Акуст. журн. 2005. Т. 51. № 4. С. 536–542.
37. *Whiteside S., Henry L., Dobbin R.* Sex differences in voice onset time: A developmental study of phonetic context effects in British English // J. Acoust. Soc. Am. 2004. V. 116. N 2. P. 1179–1183.
38. *Sussman H.M., McCaffrey H.A., Matthews S.A.* An investigation of locus equations as a source of relational invariance for stop place categorization // J. Acoust. Soc. Am. 1991. V. 90. P. 1309–1325.
39. *Iskarous Kh., Fowler C.A., Whalen D.H.* Locus equations are an acoustic expression of articulator synergy // J. Acoust. Soc. Am. 2010. V. 128. N 4. P. 2021–2032.
40. *Montgomery A., Reed P.E., Crass K.A., Hubbard H.I., Stith J.* The effects of measurement error and vowel selection on the locus equation measure of coarticulation // J. Acoust. Soc. Am. 2014. V. 136. N 5. P. 2747–2750.
41. Речь. Артикуляция и восприятие. М.: Наука, 1965.
42. *Klatt D.H.* Linguistic uses of segmental duration in English sentences // J. Acoust. Soc. Am. 1976. N 5. P. 1208–1221.
43. *Olson D.J.* Phonetic feature size in second language acquisition: Examining VOT in voiceless and voiced stops // Second Language Research. 2022. V. 38(4). P. 913–940.

Pulse Source of Excitation in Speech Signal

V. N. Sorokin

Institute for Information Transmission Problems of the Russian Academy of Sciences, B. Karetny per. 19, Moscow, 127994 Russia

e-mail: vns@iitp.ru

The properties of speech bursts of closure are studied using the material of a database of 39 speakers containing single-digit and multi-digit numerals with parallel recording of signals on a telephone handset and a directional microphone. Speech burst detection is performed by a short-term and long-term detector of spectral-temporal inhomogeneities, as well as a detector of the similarity measure of the eigenfunctions of the consonant burst spectrum and the current spectrum of the speech burst. The probability of the presence of a voiced or voiceless closure is estimated in the spaces of the amplitude spectrum and the spectrum of the group delay by the ratio of energy in the high and low frequency ranges. The place of articulation of a back-lingual consonant affects the probability distributions of the duration of the interval between the onset of a speech burst and the onset of a vowel, the frequency of the peak with maximum amplitude in the high-frequency region, the ratio of the energy in the high- and low-frequency region of the speech burst spectrum, and the similarity measures of the eigenfunctions of the consonant burst spectrum and the current spectrum of the speech burst.

Keywords: speech recognition, speech signal segmentation, stop-stop detector, plosive consonant recognition