МОДЕЛИРОВАНИЕ АУДИТОРНОЙ СУППРЕССИИ В ЧАСТОТНОЙ ОБЛАСТИ НА ОСНОВЕ СДПФ ДЛЯ ВЫДЕЛЕНИЯ ПРИЗНАКОВ РАСПОЗНАВАТЕЛЕЙ РЕЧИ ПОВЫШЕННОЙ ЭФФЕКТИВНОСТИ В УСЛОВИЯХ ШУМОВ

Иванов А.В.*, Петровский А.А.**

Белорусский Государственный Университет Информатики и Радиоэлектроники
* alexei_v_ivanov@ieee.org ** palex@it.org.by

Реферат. Статья посвящена дальнейшему исследованию и алгоритмической интерпретации явления аудиторной суппрессии, наблюдаемому на физиологическом уровне во внутреннем ухе человека. Антропоморфический алгоритм выделения признаков, основанный на модели суппрессии, отличается повышенной, по сравнению с классическими методами, эффективностью в условиях шумов. Этот факт зафиксирован на основании оценок информативности признаков, а также, их инвариантности к присутствию в составе исходного речевого сигнала шумов. Благодаря применению сжатого дискретного преобразования Фурье показана возможность кардинально снизить потребление вычислительных ресурсов по сравнению с существующими алгоритмами, учитывающими явление суппрессии.

1. Антропоморфическое моделирование

Антропоморфические, т.е. действующие подобно человеку, алгоритмы представляют собой одно из перспективных направлений развития систем распознавания речи с целью придания им большей эффективности в условиях реальной акустической обстановки. Из сравнения производительности человека и машин становится очевидным, что производительность существующих искусственных систем значительно ухудшается под влиянием шумов и помех, характерных для реальной жизни, в то время как производительность человека в схожих условиях является практически неизменной. Такая сильная подверженность искусственных систем влиянию окружающей обстановки является одним из наиболее важных факторов, ограничивающих широкое применение систем распознавания речи в составе технических устройств. В такой ситуации очевидным решением проблемы эффективности распознавателей речи является алгоритмическая интерпретация методов обработки звуковых сигналов живой системой и использование полученных алгоритмов в составе искусственных устройств, что и составляет суть антропоморфического подхода к конструированию алгоритмов.

Большинство наиболее широко применяемых методик выделения признаков, в целом, следуют тому же пути, что и слух человека. Первым шагом на пути к получению представления звукового отрывка в виде признак-вектора является получение сонограммы звукового отрывка в частотно-временной плоскости, что хорошо согласуется со “спектральным анализом” слуховой улитки, когда частота возбуждающего тонового сигнала транслируется в место максимальной амплитуды колебаний на базилярной мембране. Сонограмму традиционно получают при помощи банка полосовых фильтров с последующей децимацией, кратковременного преобразования Фурье, либо при помощи вэйвлет-анализа. В различных методиках сонограмму принято далее преобразовывать к перцептуальной шкале частот, подвергать перцептуальному интегрированию для её более полного соответствия одной из существующих психоакустических моделей слуха.

Психоакустически обоснованное моделирование слуха противоречиво. Признавая, что вцелом слух человека представляет собой нелинейный способ обработки сигналов, его свойства описываются на основе экспериментов с простыми звуками и обобщаются на случай произвольных сложных сигналов. Одной из возможных альтернатив психоакустическому моделированию, при котором ставится задача получения эмпирических соотношений между субъективными откликами слушателя и входными сигналами, является моделирование физиологическое, разбивающее исходный нелинейный “чёрный ящик” аудиторного аппарата на составные элементы и ставящее в соответствие этим элементам модели, отражающие суть объективно зафиксированных физических процессов, протекающих под влиянием входных воздействий.

Помимо уже отмеченной способности транслировать частоту возбуждающего тона в место максимальной амплитуды колебаний на базилярной мембране, оно характеризуется физиологически зафиксированным свойством аудиторной суппрессии, которая заключается в объективном снижении уровня отклика на более высокочастотные компоненты в составе сложного сигнала в присутствии низкочастотных компонентов. Аудиторная суппрессия наряду с “распространением возбуждения” в настоящий момент признаётся одним из механизмов, ответственных за психоакустически наблюдаемое явление маскирования. Явление суппрессии описывается моделями активной улитки [1], которые характеризуются сигнало-зависимой переменностью коэффициента усиления эквивалентных кохлеарных фильтров. Доводом в пользу важности аудиторной суппрессии при обработке речи ухом является тот факт, что пациенты с кохлеарными нарушениями, выражающимися в снижении эффективности “кохлеарного усилителя”, страдают от снижения способности распознавать речь на фоне шумов.

Настоящая статья ставит своей целью дальнейшее развитие идеи [1,2] моделирования наблюдаемого физиологически явления суппрессии откликов базилярной мембраны человека в составе алгоритма выделения признаков искусственной системы распознавания речи. В частности, осуществлена попытка решительного сокращения вычислительной сложности алгоритма с некоторыми упрощениями по отношению к исходной модели “кохлеарного усилителя”[1].

2. Модель аудиторной суппрессии в частотной области

Использование сжатого дискретного преобразования Фурье (англ. Warped Discrete Fourier Transform - WDFT) [3] имеет преимущества в задаче получения перцептуальной сонограммы, т. к. позволяет непосредственный переход от исходного сигнала во временной области к его представлению в частотной таким образом, что шкала частот примерно соответствует перцептуальной. Конкретная картина искажений шкалы частот задаётся конформным преобразованием z-плоскости, переводящим точки на единичной окружности в точки на единичной окружности. В частности, преобразование, соответствующее всепропускающему фильтру первого порядка:

$$z'^{-1} = A(z) = \frac{\rho + z^{-1}}{\rho z^{-1} + 1} \tag{1}$$

при $0 < \rho < 1$ позволяет растянуть отсчёты сонограммы низкочастотной и сжать высокочастотной областях:

$$tg\left(\frac{\varphi}{2}\right) = \frac{(1-\rho)}{(1+\rho)} tg\left(\frac{y}{2}\right), \tag{2}$$

где $y \in [0,\pi]$ - “линейная” шкала частот (единичная окружность в плоскости z'), а $\varphi(y) \in [0,\pi]$ - “искажённая” шкала частот, полученная в плоскости z при помощи преобразования (1). Частоты φ и y являются обобщёнными и не зависят от конкретного выбора частоты дискретизации F_s.

Физиологическая шкала частот задаётся эмпирическим соотношением, аппроксимирующим преобразование частота – место на базилярной мембране [4]:

$$f(x) = 160\left(10^{2.1x} - 0.8\right), \tag{3}$$

где $f(x) \in [F_{\min}, F_{\max}]$ - частота в Гц, соответствующая положению x на базилярной мембране, $x \in [0,1]$ - расстояние отмеренное вдоль базилярной мембраны от некоторого положения до апекса улитки. С практической точки зрения имеет смысл несколько упростить соотношение (3) заменив коэффициент 0.8 на 1. Такая замена приведёт к незначительному изменению частот, соответствующих положениям $x > 0$, и сделает точным соответствие $f(0) = 0$, что является удобным при попытке аппроксимации шкалы (3) шкалой (2). Преобразование (3) к обобщённой шкале частот приводит к следующему выражению:

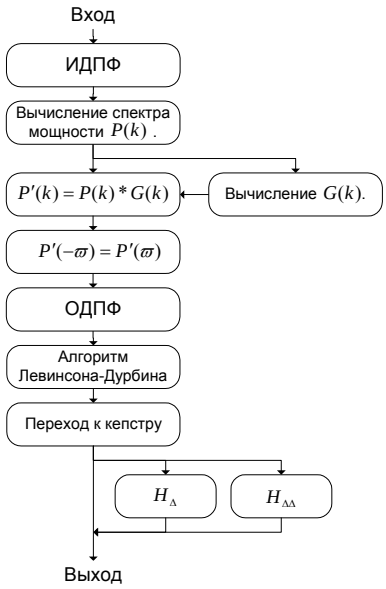

Рис. 1. Структура алгоритма выделения признаков

$$\psi = \frac{320\pi}{F_s}\left\{\left(\frac{F_s}{320}+1\right)^{\frac{y}{\pi}} - 1\right\}, \tag{4}$$

где $y \in [0,\pi]$ - “линейная” шкала частот, совпадающая с использованной в (2),а $\psi(y) \in [0,\pi]$ - идеальная перцептуальная шкала частот, аппроксимация которой желательна при помощи $\varphi(y)$.

Из сравнения (2) и (4) становится очевидным, что выбор конкретных значений ρ и F_s определяет степень подобия между этими шкалами частот. Численный эксперимент показывает, что значение $\rho \approx 0.60071382564373$ минимизирует значение среднеквадратического отклонения $\varphi(y)$ от $\psi(y)$:

$$S(\psi,\varphi) = \int_0^{\pi} (\psi(y) - \varphi(y))^2 \, dy \tag{5}$$

на сетке с шагом 0.001π при априори заданном значении $F_s = 16kHz$.

Сжатое ДПФ $X(k), k \in \mathrm{N}, 0 \le k < N$ вычисляется при помощи умножения отрезка исходного сигнала $x(k), k \in \mathrm{N}, 0 \le k < N$ на матрицу ДПФ в плоскости z', которая выражается в через переменную z при помощи замены переменных (1) [5].

Полученное при помощи СДПФ представление является искомым перцептуально-подобным представлением отрезка входного сигнала и может быть использовано для построения перцептуальной сонограммы при помощи метода, аналогичного скачущему преобразованию Фурье.

Дальнейшие этапы моделирования суппрессии (рис.1) включают в себя вычисление спектра мощности $P(k), k \in \mathrm{N}, 0 \le k \le N/2+1$, оценку коэффициентов усиления в каждом из частотных отсчётов полученной перцептуальной сонограммы $G(k)$, умножение спектра на эти коэффициенты и вычисление кепстра линейного предсказания. Алгоритм, обозначаемый при проведении экспериментов как “классический”, не содержит этапа оценки коэффициентов усиления $G(k)$.

Значения коэффициентов усиления $G(k)$ в каждом из перцептуальных каналов определяются при помощи умножения матрицы G на вектор (столбец) текущего значения спектра мощности. Матрица G представляет собой сумму матрицы, состоящей из строк, содержащих коэффициенты усиления фильтров G1 модели активной улитки [1,2] на частотах, соответствующих центральным частотам каждого из перцептуальных каналов, и единичной матрицы. В дальнейшем, результат умножения подвергается действию насыщающейся нелинейности, которая реализована при помощи сигмоидальной функции. Таким образом реализуется следующее соотношение:

$$G(k) = 1 - \frac{(G_{\max}(k) - G_{\min}(k))}{G_{\max}(k)} \left(2 * \left(1 + \exp\left(-\frac{1}{C} \left(P(k) + \sum_{i=0}^{k-1} P(i) * G_1(i) \right) \right) \right)^{-1} - 1 \right), \quad (6)$$

которое гарантирует изменение коэффициента усиления в выбранном перцептуальном канале от своего максимального значения, равного единице, до минимального, равного $G_{\min}(k)/G_{\max}(k)$, в зависимости от суммарной мощности сигнала в более низкочастотной, по сравнению с данным каналом, области (с учётом амплитудной характеристики фильтра G1). $G_{\min}(k)$ и $G_{\max}(k)$ представляют собой минимально и максимально возможные коэффициенты усиления в выбранном перцептуальном канале [1,2], C - масштабный коэффициент, определяющий связь величины суппрессии при данном уровне входного сигнала (в работе был принят равным 10^4). Описанная модель суппрессии обладает некоторыми упрощениями по сравнению с исходной. Во-первых, коэффициент при $P(k)$ должен быть, строго говоря, равным $G(k,t-1)$, т.е. коэффициентом усиления в данном канале в предыдущий момент времени. Во-вторых, в модели положено, что величина суппрессии зависит от суммарной мощности сигнала, что не совсем точно отражает суть происходящего во внутреннем ухе процесса.

3. Оценка эффективности алгоритма выделения признаков в шумах

Для оценки эффективности описываемого алгоритма выделения признаков в условиях шумов были оценены взаимная информация между источником речевого сообщения и компонентами признак-вектора, а также, средние расстояния между признаками, полученными на основе чистых и зашумлённых речевых отрывков.

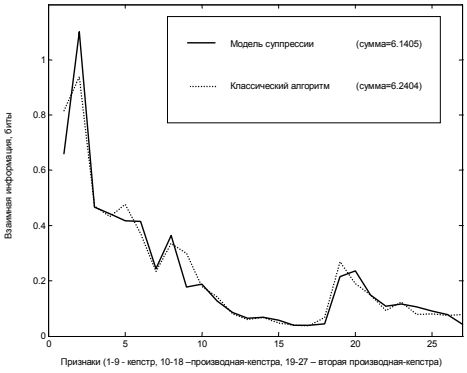

Рис. 2. Взаимная информация источника речи и компонентов признак-вектора

Эксперимент по оценке взаимной информации производился при помощи обработки фонетически размеченных чистых речевых отрывков базы TIMIT. Методика проведения эксперимента совпадает с описанной в [6] и заключается в непараметрической гистограммной оценке распределений вероятности, используемых при определении количества взаимной информации. Количество отсчётов гистограмм выбиралось на основании информационного критерия Акаике для обеспечения наилучшей аппроксимации истинного распределения вероятности, порождающего наблюдаемые признак-векторы, его гистограммной оценкой.

Результаты эксперимента, приведенные на рис. 2 подтверждают, что при моделировании суппрессии информативность признаков уменьшается в незначительной степени. В основном это связано с уменьшением информативности нулевого кепстрального коэффициента (признак №1, рис.2), которая частично компенсируется увеличением информативности кепстральных коэффициентов более высоких порядков. Вопрос о включении нулевого кепстрального коэффициента в состав вектора признаков является предметом компромисса между значительным повышением суммарного количества информации, предоставляемого вектором признаков и нежелательной зависимостью получающегося вектора от средней перцептуальной громкости входного сигнала. На основе кепстрального представления без нулевого коэффициента возможно восстанавливать сонограмму входного сигнала, но так, что его полная мощность остаётся неизменной. Это представление эквивалентно нормализации сонограммы при котором интенсивные спектральные компоненты сложного сигнала ослабляют восприимчивость к более слабым безотносительно их частотного положения, что можно принять за очень грубую модель маскирования. Очевидно, что описанная модель маскирования не позволяет увеличивать информативность кепстральных коэффициентов. Основываясь на таком рассуждении можно сделать вывод, что свойство распространения суппрессии от низкочастотных компонент к более высокочастотным является важным для возможности повысить информативность кепстральных коэффициентов более высокого, чем нулевой порядка.

Инвариантность к шумам является желательным свойством механизма выделения признаков, т.к. позволяет осуществлять тренировку статистических моделей распознавателя на незашумлённых примерах речи с возможностью их применения в условиях шумов. Методика оценки степени инвариантности также описана в [6]. Оценки Евклидова расстояния в пространстве кепстральных коэффициентов, как меры

инвариантности, производились с использованием базы речевых отрывков TI20 с добавлением белого шума с различными уровнями.

Результаты эксперимента, приведенные в табл. 1, показывают, что при помощи моделирования суппрессии становится возможным увеличение степени инвариантности кепстральных коэффициентов как в присутствии нулевого кепстрального коэффициента, так и в его отсутствии. Причём в случаях сильного шума (SNR 0-10 дБ) расстояния без нулевого кепстрального коэффициента практически соответствуют тем, что получаются при помощи классического алгоритма, но с более низким (примерно на 10 дБ) уровнем шума. При этом положительный эффект от моделирования суппрессии увеличивается с увеличением уровня шума. Проведенная в [6] оценка доверительного интервала измерения степени инвариантности однозначно указывает на статистическую состоятельность проведенного эксперимента.

Таблица 1. Средние расстояния между признак-векторами (зашумлёнными и чистыми)

Алг. / Шум	Класический Кепстр №2-9	Суппрессия Кепстр №2-9	Разница Кепстр №2-9	Классический Кепстр №1-9	Суппрессия Кепстр №1-9	Разница Кепстр №1-9
SNR=0дБ	0.6520	0.5470	0.1050	8.0333	6.8022	1.2311
SNR=10 дБ	0.5198	0.4609	0.0589	5.1899	4.9534	0.2365
SNR=20 дБ	0.4479	0.4091	0.0387	4.0526	3.9264	0.1262
SNR=30 дБ	0.2873	0.2753	0.0120	1.9997	1.9685	0.0313

4. Выводы

Описанный антропоморфический алгоритм, основанный на модели аудиторной суппрессии, является средством повышения эффективности распознавателей речи в условиях шумов поскольку позволяет увеличивать степень инвариантности признак-векторов к шумам, в частности белому шуму, эксперименты с которым проводились в работе. Моделирование суппрессии позволяет достигать измеренное увеличение инвариантности за счёт незначительной потери информативности компонентов признак-вектора.

Отличительной особенностью алгоритма является тот факт, что измеренные преимущества достигнуты при помощи упрощённой модели аудиторной суппрессии, которая предполагает небольшое увеличение количества операций, по сравнению с классическим способом выделения признаков. Дополнительный выигрыш в количестве потребляемых ресурсов достигается за счёт применения СДПФ для прямого преобразования отрезков входного звукового сигнала в сонограмму с частотной шкалой, близкой к перцептуальной.

В итоге, предлагаемый алгоритм является намного более вычислительно-эффективным средством выделения признаков в сравнении с алгоритмами, основанными на более подробной модели аудиторной суппрессии [1,2,6], обеспечивая при этом выигрыш от применения модели аудиторной суппрессии.

5. Литература

[1] Ivanov, A., Petrovsky, A. “Auditory Models for Robust Feature Extraction: Suppression”, Proc. of IEEE Signal Processing Workshop’2003, pp. 23-28, Poznan, October 10th, 2003

[2] Ivanov, A.V., Petrovsky, A. A., "A composite physiological model of the inner ear for audio coding", 116th AES Convention, preprint 6082, Berlin, Germany, May 8–11th 2004

[3] Makur, A., Mitra, S.K., “Warped Discrete-Fourier Transform: Theory and Applications”, IEEE Trans. On Circuits And Systems—I: Fundamental Theory And Applications, Vol. 48, No. 9, pp. 1086-1093, September 2001

[4] Greenwood, D. D., “A cochlear frequency-position function for several species—29 years later”, J. Acoust. Soc. Am. 87(6), pp. 2592-2605, June 1990

[5] Parfieniuk, M., Petrovsky, A., “Warped DFT as the Basis for Psychoacoustic Model”, Proc. ICASSP’04, vol. IV, pp. 185-188, May 17-21, 2004, Montreal, Canada

[6] Ivanov, A. V., Petrovsky, A. A., Ivanov, A. V., Petrovsky, A. A., “Anthropomorphic Feature Extraction Algorithm for Speech Recognition in Adverse Environments”, Proc. of the SPECOM’04 Int. Conf., pp.166-173, St. Petersburg, September 20-22, 2004