7-я Международная Конференция и Выставка Цифровая Обработка Сигналов и ее Применение DSPA-2005
Москва 16-18 марта 2005 г.

ПОМЕХОУСТОЙЧИВЫЙ ВЫДЕЛИТЕЛЬ ОСНОВНОГО ТОНА РЕЧИ

Бабкин В. В.

Центр Цифровой Обработки Сигналов, Санкт-Петербургский Государственный Университет Телекоммуникаций им. проф. М.А. Бонч-Бруевича, http://www.dsp-sut.spb.ru, http://www.dsp.sut.ru, 193382, Санкт-Петербург, пр. Большевиков д.22 корп. 1, тел. (812)589-82-43, e-mail: vb@dsp-sut.spb.ru.

Реферат. Разработанный выделитель основного тона вокализованной речи позволяет в 1,5–2 раза снизить количество грубых ошибок выделения ОТ для текущей речи, предъявляемой в белом шуме, со значением отношения сигнал/шум 0 дБ, по сравнению с существующими алгоритмами выделения ОТ, реализованными в международных стандартах низкоскоростной компрессии речи.

1. Введение

При решении задач анализа, синтеза, компрессии и распознавания речи широко используется параметрическое описание речевых сигналов, основанное на классической модели речеобразования [4]. Частота основного тона (ОТ) вокализованной речи характеризует высоту голоса и является одним из основных параметров источника голосового возбуждения речевого тракта. Задача автоматической оценки частоты ОТ и ее траектории во времени на основе анализа текущей речи является классической и активно обсуждается в мировой научной литературе многие десятилетия (см., например, обзоры [1],[2],[3]). Не смотря на огромное число предложенных алгоритмов, устойчиво работающих при анализе широкополосной речи без шума, задача практического построения помехоустойчивых выделителей ОТ надежно работающих в присутствии шума или при ограничении частотного диапазона речи, например, в телефонных каналах связи еще далека до окончательного решения.

Представленный выделитель ОТ разработан для использования в низкоскоростных вокодерах [5] и цифровых слуховых аппаратах [6] и реализован в виде модели для ПЭВМ. Проведено сравнение помехоустойчивости представленного алгоритма с существующими алгоритмами выделения ОТ, реализованными в международных стандартах низкоскоростной компрессии речи.

2. Методика оценки помехоустойчивости выделителей ОТ

Под помехоустойчивостью выделителей ОТ понимается их способность выдавать верные оценки ОТ для вокализованной речи в присутствии помех. Для оценки помехоустойчивости необходимо ввести ее количественную меру и задать условия измерений. Сравнение помехоустойчивости выделителей ОТ в силу большого числа плохо поддающихся формализации факторов, влияющих на их работу, проводится, как правило, экспериментально [2],[8]. Количественная мера помехоустойчивости выделителей ОТ в данной работе строится на основе расчета процента грубых ошибок в оценке периода ОТ от общего числа выдаваемых оценок ОТ для заданных тестовых сигналов:

$$GPE = \frac{100\%}{K}\sum_{k=1}^{K}\begin{Bmatrix}1, & если & |NPE_k| \ge \xi \\ 0, & если & |NPE_k| < \xi\end{Bmatrix};\ NPE_k = NP_k - 1;\ \ NP_k = \frac{\hat{p}_k}{p_k}, \qquad (1)$$

где K – количество измерений ОТ, NPE_k – нормированная ошибка оценки ОТ, ξ – порог разделения «грубых ошибок» и «малых отклонений» в оценке ОТ, NP_k – нормированная оценка периода ОТ, $\hat{p}_k$ – оценка периода ОТ на выходе выделителя ОТ, p_k – контрольное значение периода ОТ для k-той точки измерения ОТ, известное заранее. Необходимость нормирования оценки ОТ перед расчетом ошибки продиктована широким диапазоном возможных значений частоты ОТ (свыше трех октав). Разделение ошибок на два класса учитывает степень их влияния, например, на качество синтетической речи в низкоскоростных вокодерах. Ошибки с $/NPE_k/ < \xi$ не учитываются, т. к. они определяются способом и погрешностью измерения. Они влияют на естественность тембра и узнаваемость голоса говорящего, но не приводят к резкому ухудшению качества и разборчивости речи на выходе вокодеров, характерному для грубых ошибок [7]. Далее на рис. 2 приведены гистограммы распределения нормированных оценок ОТ NP на выходе выделителя ОТ для различных решающих правил, из которых видно, что эти распределения носят многомодальный характер. «Малые отклонения» группируются вокруг истинной оценки $NP=1$, а «грубые ошибки» в оценке ОТ связаны с переходами на гармоники и субгармоники частоты ОТ. Таким образом, порог ξ равный 0,1 хорошо разделяет эти два класса ошибок.

В качестве тестовых сигналов использовались речевые сигналы, записанные различными дикторами, мужчинами и женщинами. Использовано два типа тестовых сигналов, первые состоят из изолированных гласных звуков (набор "V"), вторые – из чтения произвольных текстов (набор "S"). Использовались широкополосные сигналы без шума (сигналы V и S) и в смеси с белым шумом с отно-

шением сигнал/шум (*SNR*) равным 0 дБ (сигналы *V0* и *S0*). Оценка *SNR* проводилась по всей длине сигналов *N*, исключая паузы, в полосе частот от 0 до 4 кГц:

$$SNR = 10\log_{10}\left[\sum_{n=0}^{N-1} s^2(n) \Big/ \sum_{n=0}^{N-1} e^2(n)\right], \tag{2}$$

где *s(n)* – сигнал без шума, *e(n)* – шумовой сигнал. Сигналы записаны с частотой дискретизации 8 кГц и разрядностью 16 бит. Траектории контрольных оценок ОТ p_k, используемых при расчете *GPE (1)*, для сигналов наборов *V* и *S* размечены вручную.

3. Структура выделителя ОТ

Общая структура построения выделителя ОТ приведена на рис. 1.

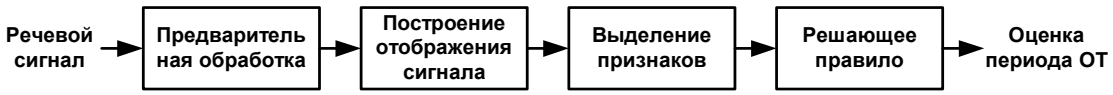
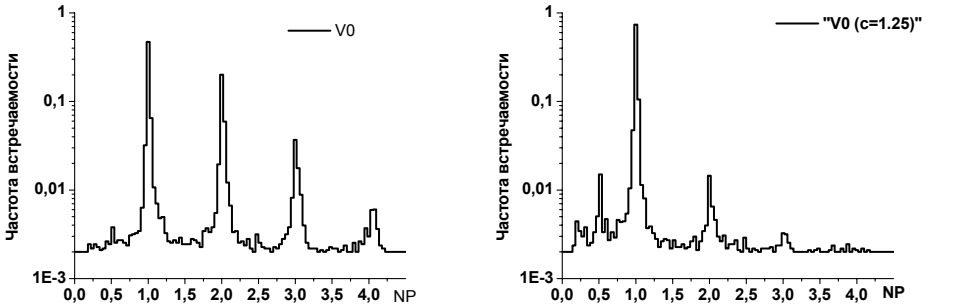


Рис. 1. Общая структура построения выделителя ОТ.

В качестве предварительной обработки сигналов используется ФНЧ с частотой среза 1 кГц. Оценка ОТ проводится на выборках длительностью 45 мс с шагом 11.25 мс. Функциональное отображение сигнала строится на основе расчета выборочной оценки функции нормированной взаимной корреляции (ФНВК) *R(p)* для сдвигов *p*, лежащих в диапазоне от $p_{min}=16$ до $p_{max}=160$:

$$R(p) = \frac{\sum_{n=1}^{N-p} x(n)x(n-p)}{\sqrt{\sum_{n=1}^{N-p} x^2(n) \sum_{n=1}^{N-p} x^2(n-p)}} \tag{3}$$

Если бы вокализованный речевой сигнал был строго периодичным, шум стационарным белым, а длина окна анализа *N* достаточно большой, тогда бы задача оценки периода ОТ сводилась бы к нахождению аргумента глобального максимума ФВНК: $p_{est}=argmaxR(p)$ в диапазоне $p_{min} \le p \le p_{max}$. Для реальных речевых сигналов такое правило порождает большое число ошибок (рис. 2-а).



а) б)

Рис. 2. Гистограммы нормированных оценок ОТ для различных решающих правил: а – выбор глобального максимума отображения *R(p)*, б – отбор кандидатов по правилу (4)

Поэтому истинная оценка периода ОТ вероятнее всего будет лежать среди аргументов положительных локальных максимумов ФНВК, которые образуют набор возможных кандидатов $\{p_m\}$. При анализе одиночного отображения обычно используются различные правила сортировки кандидатов, например, следующее:

$$for \quad p_m > p_{m-1} \quad if \quad R(p_m) > c \cdot R(p_{est}) \quad then \quad p_{est} = p_m \tag{4}$$

где *c* – порог, изменяющий баланс ошибок и минимизирующий общее их количество (рис. 2-б). Однако, количество грубых ошибок выделителя ОТ для зашумленных сигналов все еще остается существенным. Особенно много ошибок возникает в начале и в конце вокализованных участков, где степень вокализации (а следовательно и величина максимумов ФНВК) мала. Это объясняется тем, что решение принимается независимо для каждого из кадров без учета информации о возможной траектории ОТ, которая, для квазистационарных участков гласных звуков речи в первом приближении представляет собой достаточно плавную линию, характеризующую высоту голоса говорящего и общую интонацию предложения. Истинные оценки ОТ для смежных кадров для вокализованных звуков сильно коррелируют и поэтому дальнейшее снижение ошибок выделителя ОТ возможно, если учесть в решающем правиле эту зависимость. Отображение *R(p)* носит многомодальный характер, а величина его локального максимума в районе истинного периода ОТ отражает соотношение энергий перио-

дической и шумовой компонент речевого сигнала. Поэтому величины локальных максимумов $R(p_m)$ могут служить некой мерой правдоподобия оценки периода ОТ данным значением p_m при сравнении гипотез из набора $\{p_m\}$ для текущего кадра. С другой стороны, для речевых сигналов в присутствии шума выборочные оценки ФНВК испытывают значительные случайные отклонения. Поэтому для оценки наиболее вероятной траектории ОТ имеет смысл проводить совокупный анализ группы смежных кадров, учитывающий вероятности возможных кандидатов на оценку ОТ для каждого из кадров. Задача поиска траектории ОТ, таким образом, может быть сформулирована как задача динамического программирования (ДП) поиска пути, максимизирующего общую вероятность появления оценок ОТ для группы кадров с условием непрерывности траектории ОТ для гласных звуков.

В разработанном решающем правиле задача ДП сформулирована следующим образом: для речевого кадра с номером k у функционального отображения (1) ищутся аргументы всех положительных локальных максимумов, образующие набор $\{p(m_k)\}$, состоящий из M_k кандидатов на оценку периода ОТ для данного кадра. Наборы кандидатов для смежных кадров с номерами $k, k+1,..., k+K-1$ образуют столбцы решетки, через узлы которой вероятней всего пролегает траектория ОТ. Стоимость узла m_k для кадра k и кандидата $p(m_k)$ выбирается пропорциональной величине локального максимума:

$$d_N\left(m_k\right) = R_k\left(p\left(m_k\right)\right) \tag{5}$$

Стоимость пути перехода траектории ОТ между узлами решетки от кандидата $p(m_k)$ для кадра k к кандидату $p(n_{k+1})$ для кадра $k+1$ учитывает расстояние между кандидатами для смежных кадров и отражает вероятность изменения траектории ОТ, которая для вокализованных звуков считается плавной линией. В данном случае используется прямоугольная функция стоимости пути, с порогом α, задающим относительные границы поиска отклонения траектории ОТ для смежных кадров:

$$d_T\left(m_k, n_{k+1}\right) = \begin{cases} 0, & \text{если} \quad \left|p(m_k) - p(n_{k+1})\right| \le \alpha \cdot p(m_k) \\ -\infty, & \text{если} \quad \left|p(m_k) - p(n_{k+1})\right| > \alpha \cdot p(m_k) \end{cases} \tag{6}$$

Таким образом, скачки траектории ОТ, выходящие за границы интервала допустимого отклонения траектории ОТ α, не рассматриваются. Оценка наиболее вероятной траектории ОТ на протяжении K кадров осуществляется выбором оптимального пути между узлами таблицы m_k, m_{k+1},.., m_{k+K-1}, максимизирующего функционал общей стоимости пути вида:

$$D_L\left(m_k, m_{k+1}, ..., m_{k+K-1}\right) = \sum_{i=0}^{K-2}\left(d_N(m_{k+i}) + d_T\left(m_{k+i}, m_{k+i+1}\right)\right) + d_N(m_{k+K-1}), \tag{7}$$

где k – индекс кадра, m_k – индекс кандидата $p(m_k)$ на оценку периода ОТ для кадра k, K – количество смежных кадров, участвующих в сглаживании траектории ОТ.

Величина K выбирается в зависимости от допустимой задержки выдачи оценки ОТ при работе выделителя ОТ в реальном масштабе времени. В данном выделителе анализ проводится по пяти кадрам с выдачей оценки для среднего. Анализ проводится в два этапа. Сначала ищется наиболее вероятная траектория ОТ для текущего и двух будущих кадров с учетом правил (7) и (4). Затем ищется наиболее вероятная траектория ОТ для текущего и двух прошлых кадров с учетом того, что для них известны предыдущие оценки ОТ. Затем из двух оценок выбирается оценка, принадлежащая траектории с большим значением D_L.

Особенностями разработанного метода оценки траектории ОТ на основе ДП, позволяющими повысить точность оценки ОТ в шумах и снизить вычислительную сложность решения задачи поиска сглаженной траектории ОТ по сравнению с существующими решениями [7][9] являются: ограничение поиска траектории ОТ конечным числом точек; прямоугольная функция стоимости, задающая границы отклонения траектории ОТ; поиск пути, максимизирующего стоимость (7), методом последовательной оптимизации; использование комбинированного метода независимой оценки траектории ОТ по прошлым и будущим кадрам с последующим выбором лучшего результата, эффективно снижающего ошибки выделителя ОТ в начале и в конце вокализованных звуков; использование при поиске траектории ОТ оценок, найденных на прошлых кадрах; выбор наилучшей траектории ОТ по правилу (4), примененному для стоимости (7); отсутствие в решающем правиле детектора тон/шум.

4. Выбор параметров выделителя ОТ

Выбор алгоритма выделения ОТ и параметров его реализации, осуществлялся экспериментально, по критерию минимизации количества грубых ошибок для зашумленных речевых сигналов. Схема эксперимента по расчету *GPE* и оптимизации параметров приведена рис. 3. К оцениваемым параметрам относились: частота среза ФНЧ, длина N и шаг окна анализа при построении ФНВК, число срезов K в решающем правиле, пороги α и c и др. параметры.

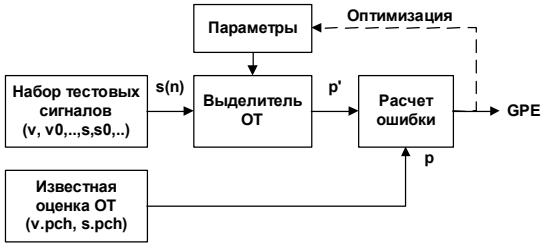


Рис. 3. Схема экспериментальной оценки помехоустойчивости выделителей ОТ.

Было проведено сравнение нескольких способов предварительной обработки сигнала, формирования отображений и построения решающих правил. В частности, в качестве функциональных отображений использовались зависимости энергии на выходе гребенчатых фильтров от частоты их настройки, а также амплитудный спектр, кепстр и автокорреляционная функция. Последние содержат практически одну и ту же информацию о сигнале, поэтому помехоустойчивость выделителей ОТ на их основе определяется, в основном, способами предварительной обработки сигнала и методами анализа отображения в решающем правиле.

5. Сравнение с существующими выделителями ОТ

Оценка помехоустойчивости разработанного выделителя ОТ и ее сравнение с помехоустойчивостью существующих выделителей ОТ велась согласно изложенной методике на основе расчета *GPE* (1) на тестовых речевых сигналах с известным значением *SNR* (2) по схеме, изображенной на рис. 3. В качестве помехи использовался белый шум. В качестве эталонных выбраны выделители ОТ, применяемые в международных стандартах низкоскоростной компрессии речи, разработанные в ходе длительных конкурсов. Наличие для них стандартных реализаций в виде Си моделей для ПЭВМ обеспечивает повторяемость полученных результатов и однозначность их трактовки. Были использованы следующие стандарты: ITU-T G.729A 8 кбит/с (1996), ITU-T G.723.1 5.3/6.3 кбит/с (1996), MIL-STD-3005 MELP 2.4 кбит/с (1998), FS-1015 LPC-10e 2.4 кбит/с (1977), ISO MPEG-4 HVXC 2.0 кбит/с (1998). Результаты сравнения помехоустойчивости выделителей ОТ представлены в таблице 1.

Таблица 1. Сравнение помехоустойчивости различных выделителей ОТ (величина *GPE* [%]).

Сигнал	G.729AB	G.723.1	MELP	LPC10E	HVXC	разработанный выделитель ОТ
V	3.9	3.1	1.7	4.8	2.4	1.4
V0	25.2	25.0	29.7	15	7.8	4.9
S	10.6	10.9	4.6	6	5.7	4.5
S0	27.5	30.7	49.3	11.1	15.9	6.9

6. Заключение

Не смотря на то, что стационарный белый шум представляет собой только узкий класс возможных помех, редко встречающийся на практике, тем не менее, он позволяет провести объективное количественное сравнение помехоустойчивости выделителей ОТ, дать понимание причин на нее влияющих и методов ее повышения. Одним из определяющих помехоустойчивость факторов является допустимая алгоритмическая задержка, позволяющая ввести адаптацию решения по будущим срезам и распознавать поведение траектории ОТ в целом.

7. Библиография

1. Hess W. Pitch determination on Speech Signals with Special Emphases on Time-Domain Methods. Proc. of NCVS Workshop on Voice Analysis, The Center of Performing Arts, Denver, February 1994.
2. Hess W. Pitch determination on Speech Signals. Springer-Verlag. New York, USA, 1983.
3. Вокодерная телефония. / Под. ред. А. А. Пирогова. М.: Связь, 1974. – 535 с., ил.
4. Рабинер Л. Р., Шафер Р. В. Цифровая Обработка Речевых Сигналов: Пер. с англ. / Под. ред. М. В. Назарова и Ю. Н. Прохорова. – М.: Радио и связь, 1981. – 496 с., ил.
5. Бабкин В.В. LPC вокодер 1000-1200 бит/с. // Труды 3-ей межд. конф. Цифровая Обработка Сигналов и ее Применение (DSPA-2000) -Москва, 2000.
6. Бабкин В.В., Бабкина Л.Н., Довжиков А.А., Молчанов А.П. Реализация карманного цифрового слухового аппарата на ADSP-2183.//2-я межд. конф. Цифровая Обработка Сигналов и ее Применение -Москва, 1999, с. 386-390.
7. Kondoz A. M. Digital Speech: Coding for Low Bit Rate Communication Systems, John Wiley & Sons Ltd, 1994.
8. Krubsack D., Niederjohn R. Comparision of Pitch Tracking Methods in Additive White Gaussian Noise. // Proc. of the 30th Midwest Symposium on Circuits and Systems, Elsevier Science Publishing Co., 1988, pp. 1262-1265.
9. Людовик Е. К. Метод определения мгновенного периода основного тона речи, основанный на динамическом программировании. Тезисы докладов и сообщений 12-го Всесоюзного семинара “Автоматическое распознавание слуховых образов”(АРСО-12). Киев-Одесса, 1982., стр. 116-119.