УДК 621.391.037.372

С. В. Алейник, М. Б. Столбов

ПОДАВЛЕНИЕ АКУСТИЧЕСКИХ ПОМЕХ АУДИОУСТРОЙСТВ С ИСПОЛЬЗОВАНИЕМ АСИНХРОННОГО ОПОРНОГО СИГНАЛА

Предложен метод двухканального шумоподавления для случая записи помехи, взятой из стороннего источника. Рассмотрены детали реализации разработанного метода, приведено сравнение его эффективности с эффективностью методов адаптивной компенсации помех.

Ключевые слова: *шумоподавление, акустические помехи, адаптивная обработка сигналов.*

Введение. Подавление помех в фонограммах является важной задачей для многих областей речевых технологий: идентификация диктора, восстановление старых фонограмм и т.п. Такая задача становится особенно актуальной, когда уровень помехи сопоставим с уровнем полезного речевого сигнала. Для ее решения предложено большое число различных алгоритмов шумоподавления [1—4].

Если помеха создается аудиоустройством и является нестационарной (пение, музыка и т.п.), эффективность одноканальных алгоритмов подавления шума уменьшается. В этом случае могут применяться двухканальные схемы адаптивной компенсации помех. В таких схемах сигнал в основном канале (основной сигнал) содержит смесь полезного речевого сигнала и помехи, а сигнал в опорном канале (опорный сигнал) содержит только помеху. Совместная обработка этих двух сигналов позволяет, при определенных условиях, эффективно подавлять помехи в основном сигнале.

Схема двухканального подавления помех представлена на рис. 1.

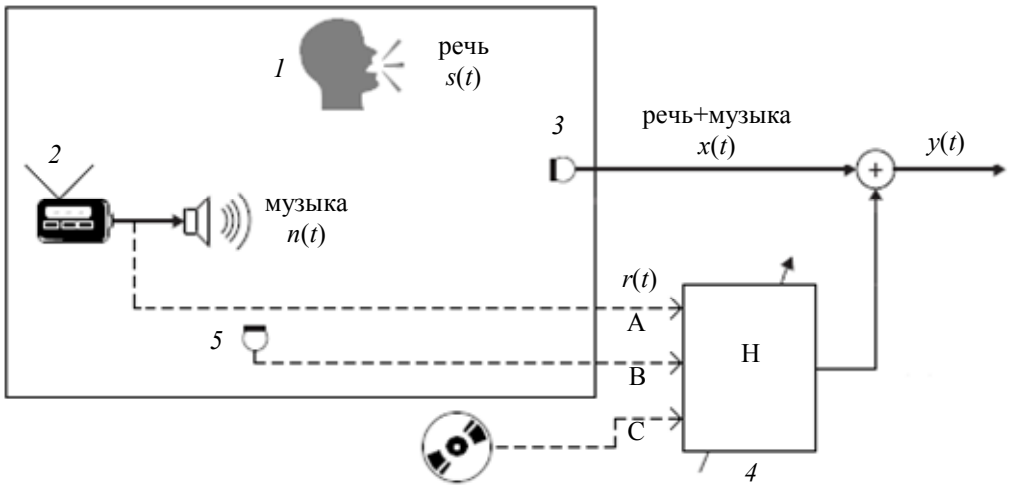

Рис. 1

Рассмотрим ситуацию записи фонограммы в помещении, когда речь $s(t)$ произносится человеком (*1*) на фоне акустической помехи $n(t)$, создаваемой работающим аудиоустройством (*2*). Речь и помеха принимаются микрофоном основного канала (*3*), формирующим основной сигнал $x(t)$. Целью обработки является подавление помехи и выделение речевого сигнала.

Процесс шумоподавления в двухканальных схемах (рис. 1) можно представить следующим образом. В дискретном случае сигналы основного и опорного каналов, $x(i)$ и $r(i)$ соответственно, описываются выражениями:

$$x(i) = h_{xs*}s(i) + h_{xn*}n(i),$$

$$r(i) = h_{rn*}n(i),$$

где i — временной индекс; $s(i)$ — речевой сигнал; $n(i)$ — помеха; * — символ свертки; h_{xs}, h_{xn} и h_{rn} — импульсные характеристики среды распространения для сигналов опорного и основного каналов.

Компенсация помехи (шумоподавление) в основном канале базируется на преобразовании опорного сигнала:

$$y(i) = x(i) + \mathrm{H}[r(i)],$$

где $y(i)$ — сигнал на выходе шумоподавителя; H — оператор преобразования опорного сигнала (рис. 1, блок *4*).

В зависимости от источника опорного сигнала возможна реализация различных алгоритмов обработки. В первом случае опорный сигнал снимается непосредственно с электрической цепи перед акустическим аудиоустройством (вход „A“ на рис. 1). В этом случае задача подавления помехи формулируется как задача эхоподавления [1, 2], в которой применяются алгоритмы адаптивной компенсации помех [2].

Во втором случае подавление помехи осуществляется с использованием опорного сигнала от микрофона, расположенного вблизи акустического источника помехи (микрофон *5*, вход „B“, рис. 1).

Наконец, особым случаем является ситуация, когда запись синхронного опорного сигнала отсутствует. Однако, если известно, какой звукоряд является помехой, то в качестве опорного сигнала может быть использована фонограмма, взятая из стороннего источника: например, музыкальная запись на компакт-диске (вход „C“ на рис. 1). В этом случае опорный сигнал является „асинхронным“, так как записан в другое время, на другой аппаратуре и в иных условиях [3].

Целью предлагаемой работы является описание практической реализации метода шумоподавления с использованием асинхронного опорного сигнала для случая акустических помех, создаваемых аудиоустройствами в помещениях.

Постановка задачи асинхронного шумоподавления. Эффективность шумоподавления зависит от целого ряда факторов: тип аудиосистемы, условия распространения звука в помещении, особенности практической реализации алгоритма шумоподавления и т.п. Из физических соображений ясно, что в асинхронном случае характеристики помех в опорном и основном каналах будут существенно различаться. Поэтому непосредственное использование асинхронной записи помехи оказывается неэффективным вследствие двух групп факторов.

1. Отсутствие синхронизации основного и опорного сигналов:

— несовпадение начала и конца помех в основном и опорном каналах;

— несовпадение частот дискретизации сигналов основного и опорного каналов.

2. Различие характеристик каналов записи сигналов основного и опорного каналов:

— различны условия записи музыки (например, запись оркестра на высококачественный CD и микрофонная запись сигнала тракта воспроизведения ТВ-приемника);

— записи выполнены в различных помещениях — различны характеристики среды (параметры реверберации и т.п.);

— различны частотные характеристики трактов записи.

Отсутствие синхронизации требует пояснения. Несовпадение начала помех в каналах связано с тем, что в асинхронном опорном сигнале помеха представляет собой полный звуко-

ряд, например, студийную запись музыкального произведения. Помеха в основном канале является только участком данного звукоряда, на который наложен полезный речевой сигнал. Начало данного участка может соответствовать любому месту музыкального произведения. Непростым является случай, когда короткий речевой сигнал начинается и заканчивается на участке, соответствующем припеву (или иному повторяющемуся фрагменту) в песне, что вызывает трудности в конкретной локализации участка помехи.

Несовпадение частот дискретизации опорного и основного сигналов также является общей проблемой для асинхронного случая. Обычно взятый с CD опорный сигнал представляет собой высококачественную запись, выполненную с частотой дискретизации 44 100 Гц. При этом основной сигнал дискретизирован с другой частотой, например, 11 025 Гц. В этом случае частота дискретизации опорного сигнала приводится к частоте основного с помощью известных алгоритмов. Однако даже после данной процедуры возможно незначительное различие в частотах дискретизации.

Такое различие приводит к тому, что в дискретизированных опорном и основном сигналах на одинаковый временной интервал приходится различное количество отсчетов. Например, в одном из случаев при анализе фонограмм частота дискретизации основного и опорного сигналов оказалась равна 16 и 16,0941 кГц соответственно, т.е. уже на десятой секунде разница в количестве отсчетов между опорным и основным сигналами составляла 941. Поскольку обработка велась покадрово, а размер кадра был выбран равным 512 отсчетам, то текущий и все последующие кадры уже не соответствовали друг другу, что привело к полной потере эффективности шумоподавления.

Различие условий записи основного и опорного сигналов в асинхронном случае также является важным фактором. Известно [2, 3], что эффективность шумоподавления адаптивных компенсаторов помех зависит от когерентности сигналов в опорном и основном каналах. Различие условий записи сигналов значительно снижает их когерентность, вследствие чего адаптивные компенсаторы оказываются малоэффективными.

Однако физические предпосылки для подавления шума в асинхронном случае все же существуют, поскольку помеха в основном и опорном каналах представляет собой различные реализации одного и того же звукоряда.

Для решения поставленной задачи нами был разработан полуавтоматический метод асинхронного шумоподавления, состоящий из двух основных шагов:

1) синхронизация основного и опорного сигналов;

2) подавление помехи в основном канале с использованием сигнала опорного канала.

Синхронизация основного и опорного сигналов представляет собой выполнение следующей последовательности действий:

— грубая синхронизация основного и опорного сигналов;

— точное совмещение начала помехи в основном и опорном сигналах;

— синхронизация частот дискретизации основного и опорного сигналов.

Грубая синхронизация выполняется оператором и включает в себя:

— приведение сигналов к единой частоте дискретизации (обычно это частота дискретизации сигнала основного канала);

— приведение средних спектров мощности сигналов к единому виду;

— приближенное определение (на слух, по спектрограмме и/или осциллограмме) начала и конца соответствующих друг другу участков помехи в опорном и основном каналах и расставление меток начала и конца участков помехи;

— приближенное совмещение участков начала помехи в опорном и основном сигналах.

Точное совмещение начала фрагментов с помехой в опорном и основном сигналах выполняется автоматически с использованием метода определения задержки сигнала по взаимокорреляционной функции [5]. Однако поскольку помехи в опорном и основном

каналах практически некоррелированы, то оценка по максимуму взаимокорреляционной функции сигналов неэффективна (максимум слабо выражен или отсутствует).

С другой стороны, кратковременные огибающие спектра мощности основного и опорного сигналов $P_x(t)$ и $P_r(t)$ на участках помехи оказываются в значительной степени коррелированными [6], так как кратковременные огибающие спектра мощности менее подвержены влиянию среды распространения и акустических трактов устройств записи—воспроизведения. Поэтому синхронизация осуществлялась по максимуму взаимной корреляции огибающих мощности опорного и основного сигналов $P_x(i)$ и $P_r(i)$:

$$P_x(i) = \left\langle x^2(i) \right\rangle \text{ и } P_r(i) = \left\langle r^2(i) \right\rangle,$$

где $\langle\ \rangle$ — символ сглаживания по времени; i — временной индекс.

С целью снижения временных затрат для оценки огибающих использовался алгоритм экспоненциального сглаживания:

$$P_x(i) = \alpha P_x(i-1) + (1-\alpha)x^2(i),$$

где $0 \le \alpha < 1$ — постоянная сглаживания, задаваемая таким образом, чтобы соответствовать темпу музыки, т.е. чтобы сигнал усреднялся без потери информации о колебаниях огибающей.

Далее, на начальных участках помехи в основном и опорном каналах (5—10 с) вычисляется взаимокорреляционная функция огибающих мощности $C(m)$:

$$C(m) = \sum_i (P_x(i) - \overline{P_x})(P_r(i-m) - \overline{P_r}),$$

где $\overline{P_x}$ и $\overline{P_r}$ — средние значения для $P_x(i)$ и $P_r(i)$ соответственно.

После этого для синхронизации начала помехи осуществляется сдвиг опорного сигнала на число отсчетов, соответствующих максимуму функции $C(m)$.

Точная синхронизация частот дискретизации также выполняется по максимуму взаимной корреляционной функции $C(m)$, вычисленной на участках, помеченных как окончание помехи в опорном и основном каналах. Если максимум $C(m)$ не соответствует нулевому сдвигу, то частоты дискретизации основного и опорного сигналов различаются. Тогда вычисляется относительный коэффициент сжатия/растяжения опорного сигнала:

$$S = (N_r + \arg\max\left(C(m)\right)\big/N_r,$$

где N_r — число отсчетов между метками начала и конца помехи в опорном сигнале. Если $S > 1$, то выполняется сжатие опорного сигнала, если $S < 1$ — то растяжение. Сжатие (растяжение) в экспериментах выполнялось на основе поотсчетной интерполяции, при этом линейная и квадратичная интерполяция давала практически одинаковые результаты.

Шумоподавление на основе метода спектрального вычитания. Опорный сигнал, полученный в результате точной синхронизации, может быть использован для компенсации помехи в основном канале. Однако применение линейных адаптивных компенсаторов в данном случае оказалось малоэффективным, что объясняется существенным уменьшением когерентности помехи в основном и опорном каналах вследствие различия условий записи и проведенных преобразований. Для этих условий наиболее подходит использование алгоритмов спектрального вычитания (АСВ) [7—9], поскольку АСВ не учитывают фазовых соотношений и позволяют подавлять помехи в случае их слабой когерентности в опорном и основном каналах.

Двухканальный АСВ организован следующим образом [1]. Мгновенный спектр Фурье на кадре основного сигнала может быть представлен в виде суммы спектров полезного речевого сигнала и спектра помехи:

$$X(f,k) = S(f,k) + N(f,k),$$

где f — частота и k — временной индекс кадра.

Спектральное вычитание определяется как [1]:

$$\left|Y(f,k)\right| = \left|X(f,k)\right| - \bar{N}(f,k),$$

где $Y(f,k)$ — оценка спектра выходного сигнала; $\bar{N}(f,k)$ — оценка амплитудного спектра помехи.

В этом случае $\left|Y(f,k)\right|$ может быть записан как:

$$\left|Y(f,k)\right| = G(f,k)\left|X(f,k)\right|,$$

где $G(f,k)$ — целевая функция фильтра шумоподавления вида:

$$G(f,k) = 1 - \bar{N}(f,k)/\left|X(f,k)\right|.$$

В более общем виде целевая функция определяется как [1]:

$$G(f,k) = \max\left[b, 1 - a\bar{N}(f,k)/\left|X(f,k)\right|\right],$$

где a и b — параметры алгоритма „коэффициент вычитания“ и „глубина подавления шума“ соответственно.

Спектр сигнала после шумоподавления рассчитывается с применением целевой функции фильтра к исходному комплексному спектру сигнала:

$$Y(f,k) = G(f,k)X(f,k).$$

Временной сигнал $y(i)$ на выходе шумоподавителя вычисляется путем обратного преобразования Фурье последовательности спектров $Y(f,k)$.

Поскольку спектр мощности шума в основном канале неизвестен, то в вычислениях используется его оценка, определяемая следующим образом. В реверберирующем помещении оценка комплексного спектра помехи может быть представлена как сумма спектров ранней и поздней реверберации [7]:

$$N(f,k) = A_0(f)R_a(f,k) + \sum_m A_m(f)R_a(f,k-m),$$

где $A_0(f)$ — фильтр, описывающий эффекты ранней реверберации; $A_m(f)$ — передаточные функции, соответствующие задержке на m кадров; $R_a(f,k)$ — комплексные спектры помехи.

Предполагая, что фазы спектров для отдельных кадров некоррелированы, мгновенный спектр мощности помехи аппроксимируем как:

$$\left|N(f,k)\right|^2 = \left|A_0(f)\right|^2\left|R_a(f,k)\right|^2 + \sum_m \left|A_m(f)\right|^2\left|R_a(f,k-m)\right|^2.$$

В рамках предлагаемого алгоритма нами учитывался только шум, порожденный ранней реверберацией, т.е.

$$N(f,k) = A_0(f)R_a(f,k).$$

В случае использования фонограммы в качестве опорного сигнала мгновенные спектры опорного сигнала $R(f,k)$ преобразуются в спектры акустической помехи путем умножения на частотный отклик $B(f)$ аудиосистемы:

$$R_a(f,k) = B(f)R(f,k).$$

Тогда спектр помехи в основном канале может быть представлен следующим соотношением:

$$\bar{N}(f,k) = A_0(f)B(f)R(f,k) = W(f,k)R(f,k),$$

где $W(f,k)$ — передаточная функция преобразования опорного сигнала. Передаточная функция может изменяться в зависимости от положения диктора и акустической обстановки в помещении, поэтому необходим адаптивный алгоритм ее оценки. В работе [9] предложен алгоритм адаптивной оценки передаточной функции в моменты присутствия акустической помехи в опорном канале и отсутствия речи диктора в основном канале. „Музыкальная“ помеха, как правило, присутствует непрерывно. При этом детектировать паузы в речи диктора представляется затруднительным ввиду нестационарного характера помехи, особенно при ее высоком уровне. Для подобного случая нами предложен следующий алгоритм оценки передаточной функции $\hat{W}(f,k)$:

$$\hat{W}(f,k) = \hat{W}(f,k-1) + \mu\left(|X(f,k)| - \hat{W}(f,k-1)|R(f,k)|\right) \Big/ \left(|X(f,k)|^2 + |R(f,k)|^2\right),$$

где $\mu < 1$ — скорость адаптации.

Экспериментальные исследования подтвердили работоспособность алгоритма оценки передаточной функции на разных типах тестовых и модельных сигналов.

С учетом оценки передаточной функции результирующий АСВ описывается следующим выражением:

$$G(f,k) = \max[b,\ 1 - a\hat{W}(f,k)|R(f,k)|/|X(f,k)|].$$

Экспериментальная оценка эффективности разработанного алгоритма. Работоспособность предложенного метода проиллюстрируем результатами следующего эксперимента. В помещении (6×5×3 м, время реверберации 480 мс) располагалась акустическая колонка. Через колонку проигрывались записанные в компьютере с частотой 16 кГц тестовые моносигналы длительностью 1,5 мин каждый: музыка, речь и розовый шум. Принятый через микрофон акустический сигнал основного канала записывался на цифровой диктофон. Микрофон основного канала в первой сессии располагался на расстоянии 1 м от акустической колонки; в последующих сессиях — на расстоянии d=2, 3 и 4 м соответственно. Одновременно тот же диктофон синхронно записывал акустический сигнал опорного канала, микрофон которого находился на расстоянии 1 м от акустической колонки. Сигналы микрофонов дискретизировались с частотой 16 кГц.

Обработка заключалась в шумоочистке сигнала основного канала с использованием различных алгоритмов шумоподавления. В качестве опорного брались как сигнал, записанный с микрофона, расположенного около колонки, так и оцифрованные исходные тестовые моносигналы (таким способом моделировалась асинхронная запись сигнала из другого источника). Для количественной характеристики уровня подавления помех использовалась характеристика „уровень подавления шума“ NR (дБ) [3]:

$$NR\ (\text{дБ}) = 1/K \sum_{k=1}^{K} 10\log_{10}(R_k),$$

где $R_k = \sum_{i=1}^{M} x_k^2(i) \Big/ \sum_{i=1}^{M} y_k^2(i)$ — уровень подавления шума на k-м кадре; K — общее количество кадров; M — размер кадра в отсчетах; $x_k(i)$ и $y_k(i)$ — входной сигнал основного ка-

нала и выходной (очищенный от шума) сигнал на k-м кадре соответственно. Усредненные результаты по всем трем видам помех (музыка, речь, шум) приведены на рис. 2.

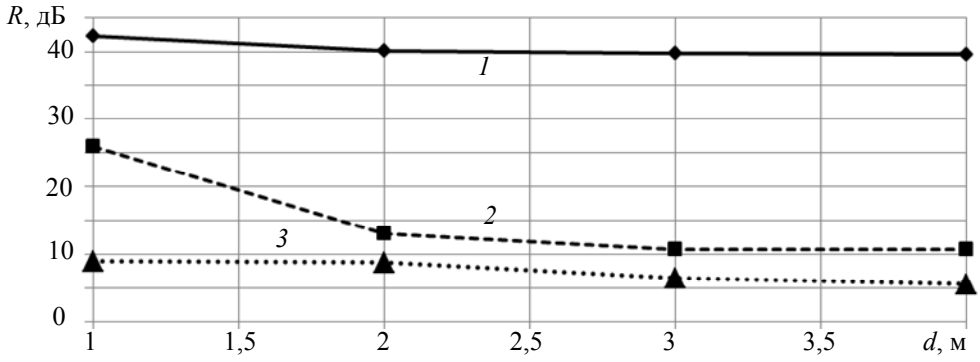

Рис. 2

Отметим, что без процедуры синхронизации как адаптивный линейный компенсатор, так и АСВ дают неудовлетворительные результаты — уровень подавления в обоих случаях практически равен нулю.

Кривая *1* подтверждает, что после синхронизации основного и опорного сигналов АСВ показывает высокую эффективность подавления помехи. С увеличением расстояния между микрофоном и излучателем степень подавления помехи при использовании АСВ снижается незначительно. Кривая *2* иллюстрирует то, что эффективность линейного компенсатора даже в случае синхронной записи помехи оказывается хуже, чем у АСВ, и значительно снижается при удалении микрофона основного канала вследствие уменьшения когерентности помех в опорном и основном каналах. Кривая *3* показывает, что применение адаптивного линейного компенсатора неэффективно в асинхронном случае.

Заключение. Предложен метод шумоподавления для записанных в помещении фонограмм, которые содержат речь, искаженную акустическими помехами, создаваемыми аудиоустройствами. Метод основан на использовании асинхронной аудиозаписи помехи, взятой из стороннего источника — CD, магнитной ленты и т.п. Метод реализуется с использованием действий, требующих участия оператора. Опыт практического применения разработанного метода для шумоочистки реальных фонограмм, поступавших от заказчиков, подтвердил его эффективность.

Центральными моментами метода являются синхронизация сигналов помехи в основном и опорном каналах и алгоритм двухканального спектрального вычитания.

В настоящее время метод встраивается в новую версию редактора Sound Cleaner, продукта ООО „ЦРТ“.

СПИСОК ЛИТЕРАТУРЫ

1. *Aalburg S., Beaugeant C., Stan S., Fingscheidt T., Balan R., Rosca J.* Single-and two-channel noise reduction for robust speech recognition in car // Siemens Corporate Research Report. Siemens AG, ICM Mobile Phones, Multimedia and Video technology, 2002.

2. *Уидроу Б., Стирнз С.* Адаптивная обработка сигналов / Пер. с англ., под ред. *В. В. Шахгильдяна.* М.: Радио и связь, 1981. 440 с.

3. *Bitzer J., Brandt M.* Speech Enhancement by Adaptive Noise Cancellation: Problems, Algorithms and Limits // AES 39th Intern. Conf. Hillerød/Dänemark, 2010. P. 106—113.

4. *Haykin S.* Adaptive Filter Theory. NY: Prentice Hall, 1996. 989 p.

5. *Benesty J., Chen J., Huang Y.* Time Delay Estimation via Linear Interpolation and Cross Correlation // IEEE Transactions on Speech and Audio Processing. 2004. Vol. 12, N 5.

6. *Ignatov P., Stolbov M., Aleinik S.* Semi-Automated Technique for Noisy Recording Enhancement Using an Independent Reference Recording // AES 46th Intern. Conf. Denver, USA, 2012.

7. *Wang L., Nakagava S., Kitaoka N.* Blind Dereverberation Based on Spectral Subtraction by Multi-channel LMS Algorithm for Distant-talking Speech Recognition // IEICE Trans. Inf. Syst. 2011. E94-D(3). P. 659—667.

8. *Бобцов А. А., Колюбин С. А., Пыркин А. А.* Алгоритм управления по выходу с компенсацией синусоидального возмущения для линейного объекта с параметрическими и структурными неопределенностями // Науч.-техн. вестн. информационных технологий, механики и оптики. 2012. № 3 (79). С. 68—72.

9. *Nasu Y., Shinoda K., Furui S.* Cross-channel spectral subtraction for meeting speech recognition // Proc. ICASSP. 2011. P. 4812—4815.

Сведения об авторах

Сергей Владимирович Алейник — ООО „ЦРТ-инновации“, Санкт-Петербург; научный сотрудник; E-mail: aleinik@speechpro.com

Михаил Борисович Столбов — канд. техн. наук; ООО „ЦРТ-инновации“, Санкт-Петербург; старший научный сотрудник; Санкт-Петербургский национальный исследовательский университет информационных технологий, кафедра речевых информационных систем; доцент; E-mail: stolbov@speechpro.com

Рекомендована кафедрой речевых информационных систем

Поступила в редакцию 22.10.12 г.