СИСТЕМЫ ОБРАБОТКИ РЕЧЕВЫХ И АКУСТИЧЕСКИХ СИГНАЛОВ

УДК 621.391.037.372

С. В. Алейник, М. Б. Столбов

ОЦЕНКА ВРЕМЕННОГО СДВИГА МЕЖДУ АУДИОСИГНАЛАМИ С ИСПОЛЬЗОВАНИЕМ ИХ ОГИБАЮЩИХ

Предложен метод оценки временного сдвига между акустическими сигналами, записанными в условиях реверберации и нелинейных искажений, базирующийся на оценке кросскорреляции временных огибающих сигналов, проведено его сравнение с другими методами оценки временного сдвига.

Ключевые слова: *временной сдвиг, временная огибающая, кросскорреляция, речевой сигнал.*

Введение. Оценка временного сдвига (ВС) между двумя сигналами (обычно называемыми „основной“ и „опорный“) важна для решения многих задач обработки аудиосигналов [1—5]: например, при оценке направления прихода сигналов, учете задержки в алгоритмах двухканальной фильтрации и др.

Большинство способов определения ВС базируется на оценке меры „близости“ сигналов друг к другу: функции кросскорреляции (ФКК) сигналов, обобщенной кросскорреляции (generalized cross-correlation, GCC), евклидова расстояния между сигналами, а также методе преобразования фазы ФКК (phase transform, PHAT) и т.п. [6—8]. Ряд факторов, таких как реверберация, увеличение расстояния между приемниками аудиосигналов, нелинейные искажения сигналов, уменьшает сходство между сигналами, что приводит к снижению стабильности оценок ВС. На рис. 1 приведены оценки ФКК (R_x) аудиосигналов, записанных в помещении при расстоянии между основным и опорным микрофонами 1 метр (кривая *1*), 2 (*2*) и 3 (*3*). Видно, что с увеличением расстояния максимум ФКК сигналов существенно снижается.

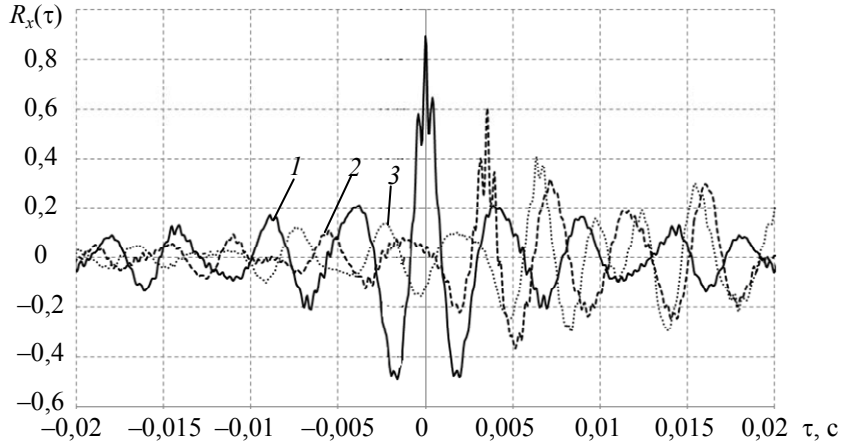

Рис. 1

В работе [9] показано, что оценка ВС на основе функции кросскорреляции временных огибающих сигналов (ФКО) дает хорошие результаты в случае сильных реверберационных искажений, назовем это методом корреляции огибающих (МКО). Обычно оценки ВС с использованием огибающих применяются в обработке коротких импульсных узкополосных сигналов в радиолокации и гидролокации [4, 7, 8, 10, 11], однако не для широкополосных аудиосигналов. Целью предлагаемой работы является описание алгоритма МКО, определение границ его применимости и оптимальных параметров.

Описание алгоритма. Оценка ВС в предлагаемом методе производится так же, как в методе ФКК. Однако сама ФКК вычисляется не по исходным сигналам, а по их временным огибающим, т.е.

$$\hat{\tau} = \arg\max(R_{a_1,a_2}(\tau)), \quad (1)$$

где $\hat{\tau}$ — оценка времени задержки, а $R_{a_1,a_2}(\tau)$ — ФКК временных огибающих основного a_1 и опорного a_2 сигналов.

Ключевым в оценке ВС (1) является вычисление огибающих. В настоящей работе для этого используется модифицированная процедура „выпрямление и фильтрация“ [12]. Обозначим дискретный временной сигнал как $x(i)$, где i — временной индекс, тогда его огибающая $a(i)$ может быть получена как:

$$a(i) = \text{ФВЧ}(\text{ФНЧ}(|x(i)|)), \quad (2)$$

где $|\cdot|$ — символ вычисления абсолютной величины (т.е. „выпрямления“) сигнала, а ФНЧ и ФВЧ — фильтры низких и высоких частот соответственно.

Фильтр низких частот предназначен для сглаживания выпрямленного сигнала и устранения выбросов. Сглаживание осуществляется фильтром первого порядка [13]:

$$y(i) = \beta(x(i) + x(i-1)) + \alpha y(i-1) \ , \quad (3)$$

где $x(i)$ и $y(i)$ — входной и выходной сигналы фильтра. Коэффициент α ($0 \le \alpha < 1$) задается на основе соотношения:

$$\alpha = 1 - 2/(1 + T_{\text{нч}} F_s), \quad (4)$$

где F_s — частота дискретизации сигнала в герцах, а $T_{\text{нч}}$ — эквивалентная длина окна в секундах, $\beta = (1-\alpha)/2$. Величина $T_{\text{нч}}$ должна соотноситься с темпом модуляции акустических сигналов (речи, музыки). Если значение $T_{\text{нч}}$ мало, то полученная огибающая будет сильно флуктуировать, если велико, это приведет к сильному сглаживанию самих огибающих. В обоих случаях уменьшится корреляция между огибающими и соответственно снизится точность оценки ВС. Таким образом, существует некая оптимальная длина окна $T^*_{\text{нч}}$.

Фильтр высоких частот предназначен для удаления постоянной и низкочастотных составляющих сглаженных огибающих. ВЧ-фильтрация также осуществляется фильтром первого порядка [13]:

$$y(i) = \eta(x(i) - x(i-1)) + \gamma y(i-1) \ , \quad (5)$$

где $\gamma = 1 - 2/(1 - T_{\text{вч}} F_s)$, а $\eta = (1+\gamma)/2$. ВЧ-фильтрация приводит, с одной стороны, к уменьшению корреляции огибающих, а с другой — к сужению главного лепестка ФКО, т.е. можно предположить, что также существует некое оптимальное $T^*_{\text{вч}}$ (заметим, что $T^*_{\text{вч}}$ и $T^*_{\text{нч}}$ в общем случае различны).

В качестве примера на рис. 2 представлены отрезок речевого сигнала (*1*), его огибающая после сглаживания (*2*) и после ВЧ фильтрации (*3*).

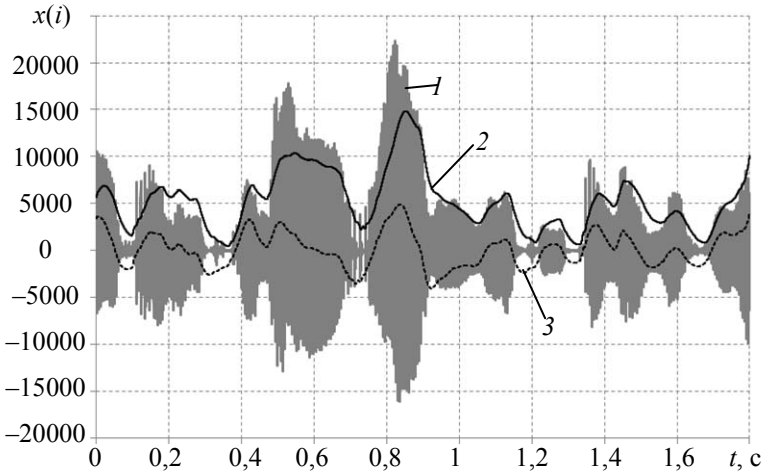

Рис. 2

Функция кросскорреляции. Качество оценки ФКО зависит от длины блока анализа данных T_a. Величина T_a должна соответствовать периодам осцилляции огибающих аудиосигнала. Если T_a < 0,1 с, то огибающая речевого сигнала может представлять собой монотонно возрастающую или убывающую функцию. В этом случае невозможно корректно оценить ФКО. Поскольку основная часть спектра огибающих расположена на частоте 4 Гц и выше, то адекватные оценки ВС получаются при $T_a \geq$ 1—2 с.

Поскольку вычисление ФКО на таких интервалах требует существенных вычислительных затрат, то вместо стандартной формулы вычисления ФКК [14]:

$$R_{x_1,x_2}(m) = \frac{\sum_i \left((x_1(i)-\bar{x}_1)(x_2(i)-\bar{x}_2)\right)}{\sqrt{\sum_i \left((x_1(i)-\bar{x}_1)^2\right)\sum_i \left((x_2(i)-\bar{x}_2)^2\right)}}$$

используем вычисление за один цикл и „с шагами“, значительно ускорив процесс без потери точности:

$$R_{x1,x2}(m) = \frac{\sum_i x_1(Ki)x_2(Ki-m) - \frac{1}{M}\left(\sum_i x_1(Ki)\sum_i x_2(Ki-m)\right)}{\sqrt{d}}, \qquad (6)$$

$$d = \left(\sum_i x_1^2(Ki) - \frac{\left(\sum_i x_1(Ki)\right)^2}{M}\right)\left(\sum_i x_2^2(Ki-m) - \frac{\left(\sum_i x_2(Ki-m)\right)^2}{M}\right). \qquad (7)$$

Здесь $K > 0$ — шаг вычисления; $x_1(i)$ и $x_2(i)$ — дискретные сигналы; N — полное число отсчетов в сигналах на блоке анализа; $m = 0, \pm 1, \pm 2, ...$ — временная задержка; $\bar{x}$ — среднее значение; $M = \lfloor (N-m)/K \rfloor$ — количество отсчетов огибающих в вычислении каждого из значений ФКО; $i = 0, ..., M-1$; $\lfloor\ \rfloor$ — символ „взятие целой части“.

Поскольку огибающая речевого сигнала осциллирует медленно, то можно задавать шаг вычисления K значительно больше единицы, что существенно ускоряет вычисления. Так как основная часть модуляционных компонент огибающих аудиосигналов находится в диапазоне до 25 Гц [15], то должно быть $K < 0{,}5F_s/25$. Для сигналов F_s=16 кГц было принято $K = 100$.

Пример ФКО реальных записей музыкальных сигналов и их огибающих представлен на рис. 3. Цифровой опорный сигнал воспроизводился через аудиоколонку. Основной сигнал был записан через микрофон в помещении с временем реверберации 650 мс, расстояние между громкоговорителем и микрофоном равнялось 4 м. Искажения основного сигнала трактом воспроизведения и реверберацией привели к тому, что корреляция между сигналами мала (кривая *1* — значение максимума, помеченное кружком, при $\tau = 0$ равно 0,11). С другой стороны, видно, что корреляция как огибающих (*2*), так и огибающих после ВЧ-фильтрации (*3*) существенна.

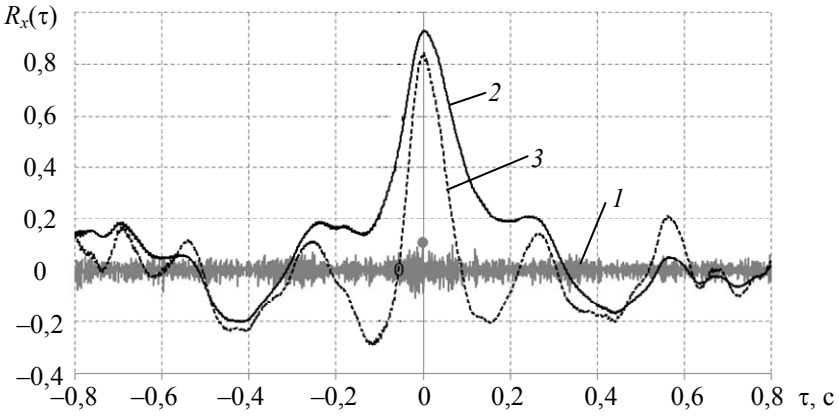

Рис. 3

Исследование влияния искажений сигналов на оценку ФКО. Пусть $x_1(i)$ и $x_2(i)$ — дискретные временные сигналы с нулевыми средними. Обозначим $R_{x_1,x_2}(m)$ — ФКК, $R_{a_1,a_2}(m)$ — ФКО сигналов. Если $x_1(i) = x_2(i)$, то $\hat{R}_{x_1,x_2}(0) = \hat{R}_{a_1,a_2}(0) = 1$ (здесь $\hat{R}$ — оценка R).

Нелинейные преобразования. Рассмотрим простые нелинейные преобразования: $x_2(i) => |x_2(i)|$, или $x_2(i) => (x_2(i))^2$. Можно показать, что в этом случае значение $R_{x_1,x_2}(0)$ существенно снижается, в то время как $R_{a_1,a_2}(0)$ меняется незначительно.

Исследование влияния шума. Зададим $x_1(i)$ и $x_2(i)$:

$$x_1(i) = (1-\mu)s(i) + \mu n_1(i), \quad x_2(i) = (1-\mu)s(i) + \mu n_2(i), \tag{8}$$

где $s(i)$ — речевой сигнал; $n_1(i)$ и $n_2(i)$ — последовательности независимых случайных величин, $0 \le \mu \le 1$. При $\mu = 0$ $x_1(i) = x_2(i) = s(i)$ и $\hat{R}_{x_1,x_2}(0) = \hat{R}_{a_1,a_2}(0) = 1$. При $\mu = 1$ $x_1(i)$ и $x_2(i)$ являются исходными независимыми случайными величинами и $\hat{R}_{x_1,x_2}(0) \approx 0$ и $\hat{R}_{a_1,a_2}(0) \approx 0$. Если дисперсии $s(i)$, $n_1(i)$ и $n_2(i)$ равны, то получим теоретические выражения для $R_{x_1,x_2}(0)$ как функцию от μ:

$$R^t_{x_1,x_2}(0,\mu) = \frac{(1-\mu)^2}{\mu^2 + (1-\mu)^2}. \tag{9}$$

На рис. 4 приведены оценки $\hat{R}_{x_1,x_2}(0)$, $\hat{R}_{a_1,a_2}(0)$, их 95 %-ные доверительные интервалы для сигналов (8) как функция от μ. Речевые сигналы брались из базы TIMIT [16], в качестве шума был взят файл factory1.wav из базы NOISEX-92 [17]. Мощности сигналов речи и шума приводились к единой величине перед преобразованием (8). Параметры вычисления огибающих: $T_a = 2$ с, $T_{\text{нч}} = 0,05$ с, ВЧ-фильтр не использовался. Полученные результаты показывают,

что при увеличении доли шума $\hat{R}_{x_1,x_2}(0)$ (кривая *1*) уменьшается, почти совпадая с теоретической кривой *3*, в то же время $\hat{R}_{a_1,a_2}(0)$ (кривая *2*) сохраняет достаточно высокие значения вплоть до $\mu = 0,6$.

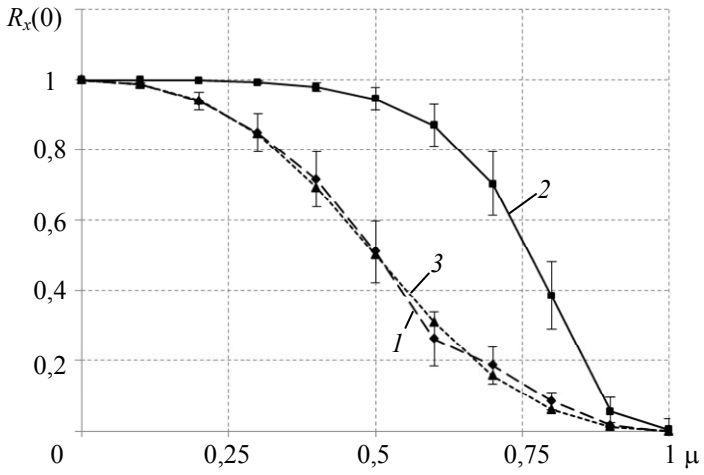

Рис. 4

Эксперименты: поиск оптимальных параметров алгоритма. Целью экспериментов являлся выбор оптимальных значений постоянных времени $T^*_{\text{нч}}$ и $T^*_{\text{вч}}$ ФНЧ и ФВЧ для различных T_a. Использовались двухканальные записи сигналов: „речь“, „песня“, „музыка“, „розовый шум“ и „модулированный по амплитуде белый шум“, записанные в помещении с постоянной времени реверберации 650 мс. Расстояние между основным и опорным микрофонами 4 м, соответственно теоретически рассчитанная задержка между сигналами для частоты дискретизации 16 кГц равнялась 183 отсчетам. В качестве целевой величины был выбран средний квадрат ошибки (mean squared error, MSE) оценки ВС:

$$\text{MSE}(\tau) = \frac{1}{L}\sum_{i=0}^{L-1}\left(\tau(i) - \tau_{\text{teor}}\right)^2,$$

где L — общее число экспериментов по оценке задержки; τ_{teor} — теоретическое значение задержки. Оптимальные значения параметров, полученные экспериментально, приведены в таблице.

T_a, с	ФНЧ	ФНЧ+ФВЧ	
	$T^*_{\text{нч}}$, с	$T^*_{\text{нч}}$, с	$T^*_{\text{вч}}$, с
2	0,0212	0,0396	0,0319
3	0,0219	0,0311	0,0441
4	0,0241	0,0313	0,0394
5	0,0107	0,0275	0,0332
6	0,0119	0,0315	0,0327
7	0,0102	0,0303	0,0275
8	0,0137	0,0225	0,0374
Среднее	0,0164	0,0321	0,0340

Сравнение МКО с другими методами оценки ВС. Предложенный метод сравнивался с кросскорреляционным и методом PHAT.

Через аудиоколонку проигрывалась музыка, записанная на компакт-диске, сигнал с которого использовался в качестве опорного, основной записывался через удаленный микрофон

в помещении и представлял собой сумму речевого сигнала и проигрываемой музыки.

Экспериментальные исследования показали, что в случаях, когда искажения основного и опорного сигналов невелики, лучшие результаты дает РНАТ (*1*), средние — ФКК (*2*), а предложенный метод (*3*) неэффективен. Однако если сигналы сильно искажены, МКО дает лучшие результаты — минимальное MSE (рис. 5).

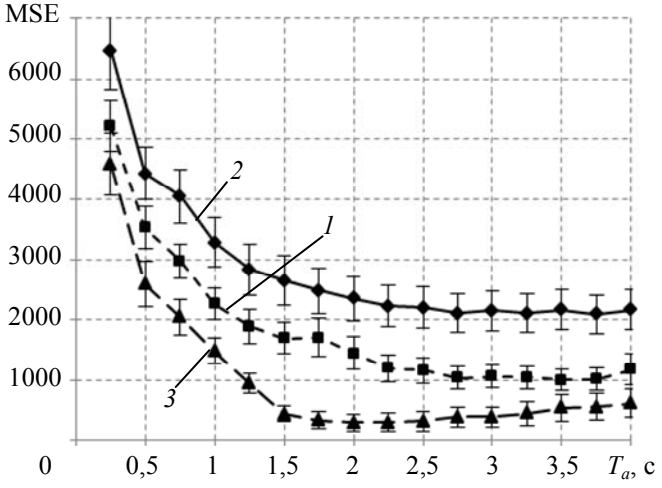

Рис. 5

Обсуждение. Полученные в работе результаты позволяют утверждать, что использование временных огибающих речевых сигналов в задаче оценки временного сдвига между аудиосигналами оправдано в случаях, когда искажения сигналов слабо влияют на огибающие. Например, МКО полезен при асинхронной фильтрации речевых сигналов [9].

Традиционные методы оценки ВС эффективнее метода МКО в случае слабых искажений самих сигналов или в случае, когда огибающие имеют сильную не меняющуюся периодичность (например, на сигналах типа „ритмичная музыка“).

По нашему мнению, вопрос выбора параметров $T_{\text{нч}}$ и $T_{\text{вч}}$ остается открытым. Эти параметры, как показывает моделирование, в общем случае зависят от характеристик как сигнала, так и его искажений. Однако соответствие полученных результатов обобщенным характеристикам спектра огибающих речевых сигналов позволяет предположить, что данные таблицы могут служить первым приближением для реальных параметров обработки.

Заключение. В работе описан и исследован метод оценки временного сдвига между двумя акустическими сигналами, основанный на кросскорреляции их огибающих. Главным достоинством метода является то, что он показывает хорошие результаты в случаях сильных искажений сигналов, например, при реверберации, или в асинхронном случае, когда сигналы записывались в разных условиях на разной аппаратуре. Недостатком является большая длина блоков данных, необходимых для оценки ВС.

Работа выполнена при государственной финансовой поддержке ведущих университетов Российской Федерации (субсидия 074-U01).

СПИСОК ЛИТЕРАТУРЫ

1. *Chen J., Benesty J., Huang Y. A.* Time Delay Estimation in Room Acoustic Environments // EURASIP J. on Advances in Signal Processing. 2006. P. 1—20.

2. *Sandmair A., Lietz M., Stefan J., Leon F. P.* Time delay estimation in the time-frequency domain based on a line detection approach // Proc. of IEEE Intern. Conf. on Acoustics, Speech and Signal Processing (ICASSP). Prague, Czech Republic, 2011. P. 2716—2719.

3. *Gedalyahu K., Eldar Y. C.* Time-delay estimation from low-rate samples: A union of subspaces approach // IEEE Transactions on Signal Processing. 2010. Vol. 58. N 6. P. 3017—3031.

4. *Kirkwood B.* Acoustic Source Localization Using Time-Delay Estimation: M.S. Thesis. Technical University of Denmark, 2003.

5. *Kozlov A., Kudashev O., Matveev Yu., Pekhovsky T., Simonchik K., Shulipa A.* SVID Speaker Recognition System for NIST SRE 2012 // Proc. of 15th Intern. Conf. "Speech and Computer" (SPECOM 2013). Springer Lecture Notes in Computer Science. Lecture Notes in Artificial Intelligence. 2013. Vol. 8113. P. 278—285.

6. *Bédard S., Champagne B., Stéphenne A.* Effects of Room Reverberation on Time-Delay Estimation Performance // Proc. of IEEE Intern. Conf. on Acoustics, Speech and Signal Processing (ICASSP). Adelaide, SA, 1994. Vol. 2. P. 261—264.

7. *Raya R., Frizera A., Ceres R., Calderón L., Rocon E.* Design and evaluation of a fast model-based algorithm for ultrasonic range measurements // Sensors and Actuators A: Physical. 2008. Vol. 148, N 1. P. 335—341.

8. *Yang L., Lavrinenko A.V., Hvam J.M., Sigmund O.* Design of one-dimensional optical pulse-shaping filters by time-domain topology optimization // Appl. Phys. Lett. 2009. Vol. 95, Is. 26. P. 261 101.

9. *Алейник С. В., Столбов М. Б.* Подавление акустических помех аудиоустройств с использованием асинхронного опорного сигнала // Изв. вузов. Приборостроение. 2013. Т. 56, № 2. С. 11—18.

10. *Lazarov B. S., Matzen R., Elesin Y.* Topology optimization of pulse shaping filters using the Hilbert transform envelope extraction // Structural and Multidisciplinary Optimization. 2011. Vol. 44, N 3. P. 409—419.

11. *Thrane N., Wismer J., Konstantin-Hansen H., Gade S.* // Application Note. Practical use of the Hilbert transform. Techn. rev. N 3. [Электронный ресурс]: <http://www.bksv.com/doc/bo0437.pdf>.

12. *Bouzid O. M., Tian G. Y., Neasham J., Sharif B.* Envelope and Wavelet Transform for Sound Localisation at Low Sampling Rates in Wireless Sensor Networks // J. of Sensors. 2012. Vol. 2012. P. 680 383.

13. *Orfanidis S. J.* Introduction to Signal Processing. [Электронный ресурс]: <http://www.ece.rutgers.edu/~orfanidi/intro2sp/orfanidis-i2sp.pdf>.

14. *Aarts R. M., Irwan R., Janssen A. J. E. M.* Efficient tracking of the cross-correlation coefficient // IEEE Transact. on Speech and Audio Processing. 2002. Vol. 10, N 6. P. 391—402.

15. *Hougast T., Steeneken H. J. M.* A review of the MTF concept in room acoustics and it's use for estimating speech intelligibility in auditoria // J. of the Acoustical Society of America. 1985. Vol. 77, Is. 3. P. 1069—1077.

16. TIMIT Acoustic-Phonetic Continuous Speech Corpus. [Электронный ресурс]: <http://catalog.ldc.upenn.edu/LDC93S1>.

17. Database of recording of various noises NOISEX-92 [Электронный ресурс]: <http://www.speech.cs.cmu.edu/comp.speech/Section1/Data/noisex.html>.

Сведения об авторах

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

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

Рекомендована кафедрой речевых информационных систем

Поступила в редакцию 22.10.13 г.