УНИВЕРСИТЕТ ИТМО
НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ
май—июнь 2019 Том 19 № 3 ISSN 2226-1494 http://ntv.itmo.ru/
SCIENTIFIC AND TECHNICAL JOURNAL OF INFORMATION TECHNOLOGIES, MECHANICS AND OPTCS
May—June 2019 Vol. 19 No 3 ISSN 2226-1494 http://ntv.itmo.ru/en/
НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК УНИВЕРСИТЕТ ИТМО ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ

УДК 004.934 doi: 10.17586/2226-1494-2019-19-3-557-559

АВТОМАТИЧЕСКОЕ РАСПОЗНАВАНИЕ РЕЧИ В УСЛОВИЯХ ШУМА МУЗЫКИ НА МНОГОКАНАЛЬНЫХ ЗАПИСЯХ С УДАЛЕННОГО МИКРОФОНА

С.С. Астапов[a], Е.В. Шуранов[b], А.В. Лаврентьев[b], В.И. Кабаров[a]

[a]Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация
[b]ООО «ЦРТ», Санкт-Петербург, 196084, Российская Федерация
Адрес для переписки: astapov@speechpro.com

Информация о статье
Поступила в редакцию 22.03.19, принята к печати 18.04.19
Язык статьи — русский

Ссылка для цитирования: Астапов С.С., Шуранов Е.В., Лаврентьев А.В., Кабаров В.И. Автоматическое распознавание речи в условиях шума музыки на многоканальных записях с удаленного микрофона // Научно-технический вестник информационных технологий, механики и оптики. 2019. Т. 19. № 3. С. 557–559. doi: 10.17586/2226-1494-2019-19-3-557-559

Аннотация

Предмет исследования. Рассмотрен метод подавления шума музыки в многоканальной записи речевого сигнала, основанный на оценке шумовой маски акустической моделью. Метод применяется для реализации автоматического распознавания речи в условиях шума музыки. **Методы.** Исследование выполнено с использованием акустической модели, реализованной на искусственных нейронных сетях, и натурных записей, сделанных в условиях реверберации. **Основные результаты.** Акустическая модель способна оценивать шумовую маску на многоканальной смеси для различных жанров музыки. Применение подобной маски для оценки ковариационной матрицы в алгоритме нацеливания MVDR (Minimum Variance Distortionless Response) способствует повышению точности распознавания речи минимум на 4,9 % на отрезке значений отношения сигнал-шум 10–30 дБ. **Практическая значимость.** Метод оценки параметров алгоритма MVDR на основе оценки шумовой маски акустической моделью способствует подавлению нестационарного шума, такого как шум музыки, что увеличивает робастность систем автоматического распознавания речи.

Ключевые слова
микрофонные решетки, MVDR, акустическая модель, оценка шумовой маски, подавление шума музыки, автоматическое распознавание речи

Благодарности
Работа выполнена при поддержке Министерства образования и науки Российской Федерации, госзадание 14.575.21.0132 (IDRFMEFI57517X0132).

doi: 10.17586/2226-1494-2019-19-3-557-559

AUTOMATIC SPEECH RECOGNITION IN PRESENCE OF MUSIC NOISE ON MULTICHANNEL FAR-FIELD RECORDINGS

S.S. Astapov[a], E.V. Shuranov[b], A.V. Lavrentyev[b], V.I. Kabarov[a]

[a]ITMO University, Saint Petersburg, 197101, Russian Federation
[b]STC Ltd., Saint Petersburg, 196084, Russian Federation
Corresponding author: astapov@speechpro.com

Article info
Received 22.03.19, accepted 18.04.19
Article in Russian

For citation: Astapov S.S., Shuranov E.V., Lavrentyev A.V., Kabarov V.I. Automatic speech recognition in presence of music noise on multichannel far-field recordings. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2019, vol. 19, no. 3, pp. 557–559 (in Russian). doi: 10.17586/2226-1494-2019-19-3-557–559

Abstract

Subject of Research. The paper considers a method of music noise reduction in a multichannel speech signal based on noise mask estimation. The method is applied for automatic speech recognition in presence of music noise. **Method.** The study is performed using an acoustic model implemented in artificial neural networks and real life recordings performed in reverberant conditions. **Main Results.** It is shown that the acoustic model is capable of estimating the noise mask on a multichannel mixture for different music genres. The application of such mask to covariance matrix estimation for MVDR (Minimum Variance

Distortionless Response) beamforming algorithm results in increasing the recognition accuracy by at least 4.9 % at signal-noise ratio levels of 10–30 dB. **Practical Relevance.** The method of MVDR coefficient estimation based on noise mask estimation by an acoustic model serves to suppress non-stationary noise, such as music, thus increasing the robustness of automatic speech recognition systems.

Keywords
microphone array, MVDR, acoustic model, noise mask estimation, music noise reduction, automatic speech recognition

Acknowledgments
This work was financially supported by the Ministry of Education and Science of the Russian Federation, Contract 14.575.21.0132 (IDRFMEFI57517X0132).

Присутствие шума музыки на фоне речи диктора является одним из самых сложных случаев для систем автоматического распознавания речи (АРР) по причине крайней нестационарности помехи. В полной мере справиться с задачей качественного подавления шума музыки неспособны как классические одноканальные методы (например, фильтр Винера), так и адаптивные многоканальные алгоритмы нацеливания, такие как алгоритм минимума дисперсии шума MVDR (Minimum Variance Distortionless Response).

Повышения качества АРР в условиях шума музыки возможно добиться, применяя акустические модели, реализованные на искусственных нейронных сетях с применением методов глубокого обучения. Для улучшения речи на многоканальных записях за последние несколько лет был предложен ряд подходов повышения робастности алгоритмов нацеливания при помощи оценивания их параметров акустическими моделями, в частности, оценка бинарной шумовой маски [1], реализация детектора речи [2], оценка ковариационной матрицы [3] и коэффициентов алгоритма нацеливания напрямую [4, 5]. В данной работе предложен метод оценки реальной шумовой маски и подсчета ковариационной матрицы для алгоритма MVDR.

Классическая реализация алгоритма MVDR осуществляется следующим образом. Для вектора смеси сигнал-шум $\mathbf{y}(t, f)$ в области кратковременного преобразования Фурье STFT (Short-Time Fourier Transform) на дискретной частоте f в момент времени t необходимо оценить сигнал диктора $\hat{x}(t, f)$. $\mathbf{y}(t, f)$ — есть вектор размерности M (количество микрофонов), а $\hat{x}(t, f)$ — скалярная величина. Алгоритм MVDR основывается на условии неискаженного приема сигнала с целевого направления θ_x и минимизации общей мощности шума. Решение задачи оптимизации приводит к следующему соотношению для вектора оптимальных весов [6]:

$$\mathbf{W}(f) = \frac{\Phi_{nn}^{-1}(f)\mathbf{h}_x(f)}{\mathbf{h}_x^H(f)\Phi_{nn}^{-1}(f)\mathbf{h}_x(f)}, \quad (1)$$

где $\mathbf{W}(f)$ — весовые коэффициенты MVDR; $\mathbf{h}_x(f)$ — вектор направленности, описывающий направление θ_x для решетки конкретной архитектуры; $\Phi_{nn}(f)$ — ковариационная матрица, посчитанная на смеси сигнал-шум; $(\cdot)^H$ — эрмитово сопряжение. Оценка сигнала выполняется с применением весовых коэффициентов (1), как $\hat{x}(t, f) = \mathbf{W}^H(f)\mathbf{y}(t, f)$. Мы предлагаем оценивать ковариационную матрицу на оценке шумовой компоненты с применением шумовой маски, оцененной акустической моделью.

Для оценки шумовой маски предложен подход, представленный в виде принципиальной диаграммы на рисунке. На основе смеси $\mathbf{y}(t, f)$ для всех частотных полос оценивается шумовая маска $\mathbf{M}_n(t, f)$, по которой в блоке COV (covariance) обновляется ковариационная матрица для MVDR $\Phi_{nn}(f)$, применяющаяся в (1): $\Phi_{nn}(f) \leftarrow \Sigma_t \tilde{\mathbf{y}}(t, f)\tilde{\mathbf{y}}^H(t, f)$, где $\tilde{\mathbf{y}}^H(t, f) = \mathbf{M}_n(t, f)\mathbf{y}(t, f)$.

Для обучения глубокой нейронной сети DNN (Deep Neural Network) использовалась база данных многоканальных записей (с 66 микрофонов) речи дикторов на фоне музыки общей продолжительностью 48 часов. Из нее 40 часов использовались для обучения и 8 — для проверки эффективности метода. В базе содержалось 15 жанров музыки. Записи производились на 66-канальную микрофонную решетку в помещении с уровнем реверберации до T_{60} = 700 мс на расстоянии 2 м до диктора. Запись музыки и запись речи дикторов производились отдельно друг от друга в идентичных акустических условиях. Смешивание производилось искусственно на заданных значениях отношения сигнал-шум (ОСШ) 10–30 дБ. Длина окна STFT составляла 512 отсчетов со сдвигом 256 отсчетов. Структура DNN для оценки шумовой маски: первые два скрытых слоя типа BLSTM (Bidirectional Long Short-Term Memory) с количеством ячеек 256 и 600 соответственно и функцией нелинейности ReLU (Rectified Linear Unit); 3 скрытый слой полносвязный (Fully Connected) с функцией нелинейности tanh и количеством ячеек 256. На вход нейронной сети подается одноканальный

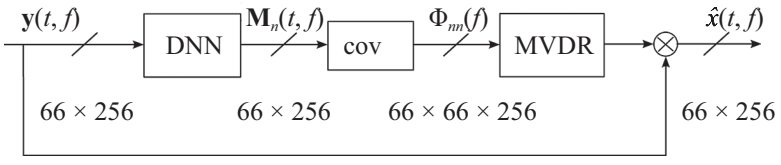

Рисунок. Принципиальная диаграмма метода очистки шума музыки

частотный вектор 1 × 256; в рабочем режиме получение маски производится путем последовательной подачи в сеть 66 частотных векторов со всех каналов записи.

Проверка метода проводилась по метрике пословной точности распознавания ACC (Accuracy). Результаты тестирования представлены в таблице. Значения ACC_0 соответствуют точности распознавания без применения алгоритма очистки шума музыки; ACC_e — точности распознавания с применением алгоритма очистки шума музыки. Результаты показывают, что представленный метод очистки шума позволяет добиться существенных улучшений точности распознавания при уровне ОСШ 10–30 дБ от 34 до 4,9 %, соответственно.

Таблица. Результаты тестирования по метрике точности распознавания речи

ОСШ (дБ)	ACC0 (%)	ACCe (%)	Прирост ACC (%)
10	37,58	71,80	34,22
15	52,92	71,84	18,92
20	61,63	72,40	10,77
30	67,84	72,74	4,90

Литература

1. Heymann J., Drude L., Haeb-Umbach R. Neural network based spectral mask estimation for acoustic beamforming // Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing. Shanghai, China, 2016. P. 196–200. doi: 10.1109/icassp.2016.7471664
2. Higuchi T., Ito N., Yoshioka T., Nakatani T. Robust MVDR beamforming using time-frequency masks for online/offline ASR in noise // Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing. Shanghai, China, 2016. P. 5210–5214. doi: 10.1109/icassp.2016.7472671
3. Li B., Sainath T.N., Weiss R.J., Wilson K.W., Bacchiani M. Neural network adaptive beamforming for robust multichannel speech recognition // Proc. INTERSPEECH. 2016. P. 1976–1980. doi: 10.21437/interspeech.2016-173
4. Yoshioka T. et al. The NTT CHiME-3 system: advances in speech enhancement and recognition for mobile multi-microphone devices // Proc. IEEE Workshop on Automatic Speech Recognition and Understanding. Scottsdale, USA, 2015. P. 436–443. doi: 10.1109/asru.2015.7404828
5. Du J. et al. The USTC-iFlyteck system for the $CHiME_4$ challenge // Proc. 4th Int. Workshop on Processing in Everyday Environments. 2016.
6. Brandstein M., Ward D. Microphone Arrays: Signal Processing Techniques and Applications. Springer, 2001. 398 p.

References

1. Heymann J., Drude L., Haeb-Umbach R. Neural network based spectral mask estimation for acoustic beamforming. *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*. Shanghai, China, 2016, pp. 196–200. doi: 10.1109/icassp.2016.7471664
2. Higuchi T., Ito N., Yoshioka T., Nakatani T. Robust MVDR beamforming using time-frequency masks for online/offline ASR in noise. *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*. Shanghai, China, 2016, pp. 5210–5214. doi: 10.1109/icassp.2016.7472671
3. Li B., Sainath T.N., Weiss R.J., Wilson K.W., Bacchiani M. Neural network adaptive beamforming for robust multichannel speech recognition. *Proc. INTERSPEECH*, 2016, pp. 1976–1980. doi: 10.21437/interspeech.2016-173
4. Yoshioka T. et al. The NTT CHiME-3 system: advances in speech enhancement and recognition for mobile multi-microphone devices. *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding*. Scottsdale, USA, 2015, pp. 436–443. doi: 10.1109/asru.2015.7404828
5. Du J. et al. The USTC-iFlyteck system for the $CHiME_4$ challenge. *Proc. 4th Int. Workshop on Processing in Everyday Environments*, 2016.
6. Brandstein M., Ward D. *Microphone Arrays: Signal Processing Techniques and Applications*. Springer, 2001, 398 p.

Авторы

Астапов Сергей Сергеевич — PhD, старший научный сотрудник, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, Scopus ID: 55320270600, ORCID ID: 0000-0001-8381-8841, astapov@speechpro.com

Шуранов Евгений Витальевич — кандидат технических наук, руководитель группы, ООО «ЦРТ», Санкт Петербург, 196084, Российская Федерация, Scopus ID: 57190970283, ORCID ID: 0000-0003-0977-5075, E_v_shuranov@mail.ru

Лаврентьев Александр Валерьевич — программист, ООО «ЦРТ», Санкт Петербург, 196084, Российская Федерация, Scopus ID: 57190961330, ORCID ID: 0000-0001-9443-2207, lavrentyev@speechpro.com

Кабаров Владимир Иосифович — старший преподаватель, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, ORCID ID: 0000-0001-6300-9473, kabarov_vi@mail.ru

Authors

Sergei S. Astapov — PhD, Senior researcher, Senior researcher, ITMO University, Saint Petersburg, 197101, Russian Federation, Scopus ID: 55320270600, ORCID ID: 0000-0001-8381-8841, astapov@speechpro.com

Evgeniy V. Shuranov — PhD, Group supervisor, STC Ltd., Saint Petersburg, 196084, Russian Federation, Scopus ID: 57190970283, ORCID ID: 0000-0003-0977-5075, E_v_shuranov@mail.ru

Alexander V. Lavrentyev — Software engineer, STC Ltd., Saint Petersburg, 196084, Russian Federation, Scopus ID: 57190961330, ORCID ID: 0000-0001-9443-2207, lavrentyev@speechpro.com

Vladimir I. Kabarov — Senior lecturer, ITMO University, Saint Petersburg, 197101, Russian Federation, ORCID ID: 0000-0001-6300-9473, kabarov_vi@mail.ru