УНИВЕРСИТЕТ ИТМО
НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ
сентябрь–октябрь 2019 Том 19 № 5 ISSN 2226-1494 http://ntv.itmo.ru/
SCIENTIFIC AND TECHNICAL JOURNAL OF INFORMATION TECHNOLOGIES, MECHANICS AND OPTCS
September–October 2019 Vol. 19 No 5 ISSN 2226-1494 http://ntv.itmo.ru/en/
НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ

УДК 004.934.5 doi: 10.17586/2226-1494-2019-19-5-951-954

АКУСТИЧЕСКОЕ МОДЕЛИРОВАНИЕ ДЛЯ СИНТЕЗА КАЗАХСКОЙ РЕЧИ

А.К. Калиев, С.В. Рыбин

Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация
Адрес для переписки: kaliyev.arman@yandex.kz

Информация о статье
Поступила в редакцию 27.06.19, принята к печати 22.07.19
Язык статьи — русский

Ссылка для цитирования: Калиев А.К., Рыбин С.В. Акустическое моделирование для снтеза казахской речи // Научно-технический вестник информационных технологий, механики и оптики. 2019. Т. 19. № 5. С. 951–954. doi: 10.17586/2226-1494-2019-19-5-951-954

Аннотация
Представлена новая конструкция генеративно-состязательной сети для обучения акустической модели синтеза речи. Предлагаемая конструкция состоит из генератора и двух дискриминаторов, где генератор предсказывает акустические параметры из лингвистического представления. Обучение и тестирование производились на корпусе казахского языка, который состоял из 5,6 ч записи речи. По результатам экспериментов была получена 3,46 средняя экспертная оценка, что говорит о достаточно приемлемом качестве синтезе речи. Данный подход может быть применим при создании технологий синтеза речи для других языков.

Ключевые слова
акустическая модель, синтез речи, казахский язык, генеративно-состязательная сеть (ГСС), речевой корпус

Благодарности
Исследования выполнены за счет стартового финансирования Университета ИТМО в рамках НИР № 618278 «Синтез эмоциональной речи на основе генеративных состязательных сетей».

doi: 10.17586/2226-1494-2019-19-5-951-954

ACOUSTIC MODELING FOR KAZAKH SPEECH SYNTHESIS

A.K. Kaliyev, S.V. Rybin

ITMO University, Saint Petersburg, 197101, Russian Federation
Corresponding author: kaliyev.arman@yandex.kz

Article info
Received 27.06.19, accepted 22.07.19
Article in Russian

For citation: Kaliyev A.K., Rybin S.V. Acoustic modeling for Kazakh speech synthesis. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2019, vol. 19, no. 5, pp. 951–954 (in Russian). doi: 10.17586/2226-1494-2019-19-5-951-954

Abstract
We present a new framework of generative adversarial network for training of acoustic model for speech synthesis. The proposed generative adversarial network consists of a generator and a pair of agent discriminators, where the generator predicts the acoustic features from the linguistic representation. Training and testing were carried out on the Kazakh speech corpus, which consisted of 5.6 hours of speech recording. According to the experiment results the 3.46 mean opinion score was obtained which shows an acceptable quality of speech synthesis. This approach of the acoustic model development can be applied in speech synthesis systems of the other languages.

Keywords
acoustic model, speech synthesis, Kazakh language, generative adversarial network (GAN), speech corpus

Acknowledgements
This work was financially supported by the initial funding from ITMO University within the framework of research practice No. 618278 “Emotional speech synthesis based on generative adversarial networks”.

В синтезе речи, как и в распознавания речи, изначально строится сложная статистическая модель, которая описывает произнесение этого звука в речи. Акустическая модель в синтезе речи позволяет генерировать акустические параметры для каждого речевого сегмента. В данной работе представлена новая конструкция генеративно-состязательной сети (ГСС) для обучения акустической модели синтезатора речи.

Как показано на рисунке, ГСС состоит из двух дискриминаторов и одного генератора. Оба дискриминатора помогают генератору учитывать распределение акустических параметров и таким образом нивелировать эффект сглаженности речевого сигнала — одной из сложнейших нерешенных проблем технологии синтеза речи [1,2]. Эффект сглаженности речевого сигнала появляется при обучении нейронной сети (НС) с использованием функции среднеквадратичного отклонения. При таком подходе перестают учитываться небольшие флуктуации в угоду уменьшения общего отклонения предсказанных данных. Однако слух человека способен воспринимать и различать все мельчайшие детали в речи, и в итоге такая речь человеком воспринимается как искусственная, или «сглаженная». Входными параметрами дискриминаторов служат как акустические, так и лингвистические параметры, благодаря этому дискриминаторы обращают внимание не только на распределение входных параметров, но и на связь лингвистических параметров с акустическими.

Была проведена Mean Opinion Score (MOS, средняя экспертная оценка) оценка качества предложенной модели ГСС на корпусе казахского языка «Асель» [3]. Выбор языка обусловлен проработанностью авторами подходов предсказания просодических параметров и других наработок в этой области [4–6]. Речевой корпус состоял из 5,6 ч нейтральной речи женского голоса или 6 тысяч отдельных фраз и предложений. Для тестового набора было случайным образом выбрано 50 предложений, и таким же образом было выбрано 50 предложений для проверочного набора. Во время проведения экспериментов акустические параметры извлекались с частотой 200 Гц (5 мс) из звукового сигнала частоты дискретизации 22 кГц. Для каждого отсчета вычислялось 97 лингвистических параметров. Акустические параметры извлекались с помощью вокодера WORLD [7]. Лингвистическими параметрами служили индексы слов и фонем, места пауз, длительность фонем и пауз, позиция фонемы в слове и в предложении, фонетические признаки фонем и другие признаки. Акустическими параметрами были F0 и мел-частотные кепстральные коэффициенты.

Традиционно ГСС состоят из конкурирующих нейронных сетей, которые условно разделяют на генератор G и дискриминатор D. Генератор предсказывает из вектора лингвистических параметров x вектор акустических параметров $\hat{y}$:$G(x)$:$x \rightarrow \hat{y}$. В то же время в дискриминатор D подается вектор акустических параметров $\hat{y}$, сгенерированный с помощью генератора G, и вектор акустических параметров y, полученный с помощью вокодера WORLD из корпуса речевых данных «Асель». Во время обучения дискриминатор учится определять, какие акустические параметры получены из реального речевого сигнала, а какие — «не настоящие». Соответственно, генератор обучается обманывать дискриминатор, предсказывая акустические параметры, максимально близкие к разметке корпуса «Асель».

На практике во время такой схемы обучения, к сожалению, генератор стремится генерировать распределение акустических параметров не близкое к естественному, а то, которое «обмануло» бы дискриминатор [2]. Для решения этой проблемы генератор предварительно обучается с использованием функции среднеквадратического отклонения (MSE) $L_{mse}(y, \hat{y})$, затем во время обучения всей ГСС проводится только несколько итераций. Таким образом, удается сохранить свойства предварительно обученного генератора, при этом полученное акустическое распределение становится более близким к естественному.

Для того чтобы дискриминатор также провоцировал генератор находить связь между акустическими и лингвистическими параметрами, предлагается подавать в дискриминатор не только вектор акустических параметров, но и вектор лингвистических параметров. Однако загружая дискриминатор дополнительными параметрами в свою очередь можно столкнуться с тем, что во время обучения дискриминатор будет невольно больше фокусироваться на лингвистических параметрах или же дискриминатору будет сложно найти взаимосвязь между лингвистическими и акустическими параметрами. Так как лингвистических параметров намного больше, чем акустических, то обучение ГСС может пойти по нежелательному сценарию. Для исключения этой ситуации было предложено расширить стандартный ГСС с «Пара-Агент» структурой, как показано на рисунке, где добавлен второй дискриминатор, принимающий на порядок меньшее количество лингвистических параметров.

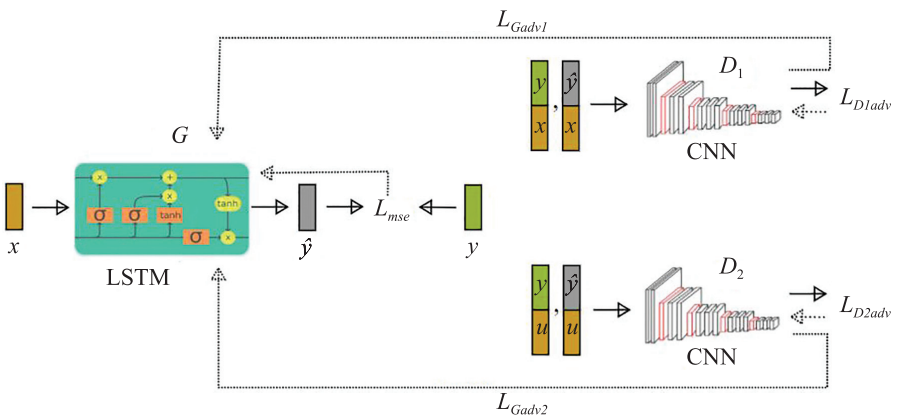

Рисунок. Архитектурная конструкция предложенной генеративно-состязательной сети

Агенты 1 и 2 — это условные дискриминаторы D_1 и D_2 соответственно, реализованные сверточными нейронными сетями (CNN). Входными данными D_1 служит вектор из акустических и лингвистических параметров. Дискриминатор D_2 принимает вектор акустических и вектор $u \subset x$ 11 наиболее важных лингвистических параметров. Эти лингвистические параметры были отобраны авторами в результате экспериментов.

Конечная функция ошибки генератора будет выглядеть следующим образом:

$$L_G = L_{mse} + w_1 \frac{E_{L_{mse}}}{E_{L_{G\,adv1}}} L_{G\,adv1} + w_2 \frac{E_{L_{mse}}}{E_{L_{G\,adv2}}} L_{G\,adv2},$$

где $L_{mse} = \left\| y - \hat{y} \right\|_2^2$ — функция среднеквадратического отклонения; L_{Gadv1} и L_{Gadv2} — соревновательная функция ошибки генератора для дискриминаторов D_1 и D_2; $E_{L_{mse}}$, $E_{L_{GadvI}}$, $E_{G_{adv2}}$ — ожидаемые значения $E_{L_{mse}}$, $E_{L_{Gadv1}}$, и $E_{L_{Gadv2}}$ соответственно; $\frac{E_{L_{mse}}}{E_{L_{G\,adv1}}}$ и $\frac{E_{L_{mse}}}{E_{L_{G\,adv2}}}$ — шкала нормализации; w_1 и w_2 — вес функций ошибок.

В качестве генератора была использована предварительно обученная нейронная сеть LSTM (Long-Short Term Memory, долгая краткосрочная память), дискриминаторами служили CNN. Архитектура LSTM — однонаправленная, 3-слойная НС, в каждом слое по 256 блоков памяти. Для каждого слоя в LSTM использовалась Tanh функция активации. Каждый дискриминатор состоял из одного сверточного слоя с пулами, ReLU функцией активации и полносвязным слоями, за которыми следует еще один полносвязный слой. На выходе получается значение, представляющее вероятность того, что входные данные являются «реальными».

Для оценки качества сгенерированной речи была проведена MOS оценка. Всего в опросе участвовали 11 носителей казахского языка. Им было предложено оценить 6 наборов записей. Для каждого набора слушателям предлагалось прослушать и оценить каждую запись отдельно. Каждую аудиозапись разрешалось прослушать неограниченное количество раз, но было рекомендовано ограничиваться 2–3 прослушиваниями.

По результатам опроса была получена MOS-оценка 3,46, что говорит о достаточно приемлемом качестве синтезе речи. По мнению авторов, для улучшения качества синтезируемой речи и приведению к результатам современного уровня достаточно увеличения выборки обучения в 3–4 раза и повышения качества разметки речевых данных. Также есть уверенность, что представленный метод может быть применим для высокоресурсных языков с большими данными обучения.

Таким образом, представлена новая конструкция ГСС для обучения акустической модели синтеза речи. На основании полученной MOS-оценки можно утверждать, что представленный подход позволяет разрабатывать технологию синтеза речи достаточно приемлемого качества для малоресурсных языков, которым является казахский язык [8]. Важно отметить, что в новой конструкции ГСС дискриминаторы обращают внимание не только на распределение акустических параметров, но и на связь лингвистических параметров с акустическими. Несмотря на то что обучение и тестирование проводилось на корпусе казахского языка, авторы предполагают, что такой подход приемлем и для других языков.

Литература

1. Ze H., Senior A., Schuster M. Statistical parametric speech synthesis using deep neural networks // Proc. IEEE International Conference on Acoustics, Speech and Signal Processing. ICASSP. 2013. P. 7962–7966. doi: 10.1109/ICASSP.2013.6639215
2. Saito Y., Takamichi S., Saruwatari H. Statistical parametric speech synthesis incorporating generative adversarial networks // IEEE/ACM Transactions on Audio, Speech, and Language Processing. 2017. V. 26. N 1. P. 84–96. doi: 10.1109/TASLP.2017.2761547
3. Khomitsevich O., Mendelev V., Tomashenko N., Rybin S., Medennikov I., Kudubayeva S. A bilingual Kazakh-Russian system for automatic speech recognition and synthesis // Lecture Notes in Computer Science. 2015. V. 9319. P. 25–33. doi: 10.1007/978-3-319-23132-7_3
4. Kaliyev A., Rybin S.V., Matveev Y. The pausing method based on brown clustering and word embedding // Lecture Notes in Computer Science. 2017. V. 10458. P. 741–747. doi: 10.1007/978-3-319-66429-3_74
5. Kaliyev A., Rybin S.V., Matveev Yu.N., Kaziyeva N., Burambayeva N. Modeling pause for the synthesis of Kazakh speech // Proc. 4th International Conference on Engineering and MIS, ICEMIS. 2018. P. 1–4. doi: 10.1145/3234698.3234699
6. Kaliyev A., Rybin S.V., Matveev Y.N. Phoneme duration prediction for Kazakh language // Lecture Notes in Computer Science. 2018. V. 11096. P. 274–280. doi: 10.1007/978-3-319-99579-3_29

References

1. Ze H., Senior A., Schuster M. Statistical parametric speech synthesis using deep neural networks. *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing. ICASSP*, 2013, pp. 7962–7966. doi: 10.1109/ICASSP.2013.6639215
2. Saito Y., Takamichi S., Saruwatari H. Statistical parametric speech synthesis incorporating generative adversarial networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2017, vol. 26, no. 1, pp. 84–96. doi: 10.1109/TASLP.2017.2761547
3. Khomitsevich O., Mendelev V., Tomashenko N., Rybin S., Medennikov I., Kudubayeva S. A bilingual Kazakh-Russian system for automatic speech recognition and synthesis. *Lecture Notes in Computer Science*, 2015, vol. 9319, pp. 25–33. doi: 10.1007/978-3-319-23132-7_3
4. Kaliyev A., Rybin S.V., Matveev Y. The pausing method based on brown clustering and word embedding. *Lecture Notes in Computer Science*, 2017, vol. 10458, pp. 741–747. doi: 10.1007/978-3-319-66429-3_74
5. Kaliyev A., Rybin S.V., Matveev Yu.N., Kaziyeva N., Burambayeva N. Modeling pause for the synthesis of Kazakh speech. *Proc. 4th International Conference on Engineering and MIS, ICEMIS*, 2018, pp. 1–4. doi: 10.1145/3234698.3234699
6. Kaliyev A., Rybin S.V., Matveev Y.N. Phoneme duration prediction for Kazakh language. *Lecture Notes in Computer Science*, 2018, vol. 11096, pp. 274–280. doi: 10.1007/978-3-319-99579-3_29

7. Morise M., Yokomori F., Ozawa K. WORLD: a vocoder-based high-quality speech synthesis system for real-time applications // IEICE Transactions on Information and Systems. 2016. V. E99-D. N 7. P. 1877–1884. doi: 10.1587/transinf.2015EDP7457
8. Карпов А.А., Верходанова В.О. Речевые технологии для малоресурсных языков мира // Вопросы языкознания. 2015. № 2. С. 117–135.

Авторы

Калиев Арман Куанышевич — аспирант, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, Scopus ID: 55701609000, ORCID ID: 0000-0001-8399-8379, kaliyev.arman@yandex.kz
Рыбин Сергей Витальевич — кандидат физико-математических наук, доцент, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, Scopus ID: 57112217900, ORCID ID: 0000-0002-9095-3168, svrybin@itmo.ru

7. Morise M., Yokomori F., Ozawa K. WORLD: a vocoder-based high-quality speech synthesis system for real-time applications. *IEICE Transactions on Information and Systems*, 2016, vol. E99-D, no. 7, pp. 1877–1884. doi: 10.1587/transinf.2015EDP7457
8. Karpov A.A., Verkhodanova V.O. Speech technologies for under-resourced languages of the world. *Voprosy jazykoznanija*, 2015, no. 2, pp. 117–135. (in Russian)

Authors

Arman K. Kaliyev — postgraduate, ITMO University, Saint Petersburg, 197101, Russian Federation, Scopus ID: 55701609000, ORCID ID: 0000-0001-8399-8379, kaliyev.arman@yandex.kz

Sergey V. Rybin — PhD, Associate Professor, ITMO University, Saint Petersburg, 197101, Russian Federation, Scopus ID: 57112217900, ORCID ID: 0000-0002-9095-3168, svrybin@itmo.ru