соответственно, в частотном канале *f*. А $p(\vec{ss}_{tf} \mid \lambda_{m=1}^{Bayes}(f))$ и $p(\vec{ss}_{tf} \mid \lambda_{m=0}^{Bayes}(f))$ соответствующие плотности вероятности надежности и ненадёжности для канала *f*, основанные на признаках $ss_{tf.}$

В качестве признаков используют различные параметры: в работе [9] – спектральная энергия полосы и её производные, в работе [8] были использованы: *коэффициент гребенчатого фильтра* для сравнения энергии в вокализованных областях с энергией в негармонических областях; *коэффициент автокорреляции*, для измерения периодичности сигнала; *отношение энергии поддиапазона к энергии полной полосы*, представляющее спектральную форму; *оценка энергии шума*; *коэффициент эксцесса*, используемый для измерения «остроты» пика сигнала; *коэффициент тональности*, для измерения SNR.

Используя предполагаемые маски, отдельные классификаторы обучаются для невокализованных и вокализованных типов внутри каждого канала.В оценках для шумоустойчивого распознавания оценки классификатора превосходят традиционные оценки спектрального вычитания во всех условиях шума, но особенно для нестационарных случаев [10].

Таким образом, методы, основанные на слуховой системе человека, так и основанные на классификации речевых параметров могут обеспечить лучшую эффективность распознавания в сравнении с методами на основе оценки только SNR. Преимуществом аудиторных подходов в оценке маски является их способность строить решения, основанные на свойствах спектра речевого сигнала. Это позволяет более точно идентифицировать доминирующие речевые спектральные области по сравнению с подходами на основе SNR, которые предполагают обобщение шумовых характеристик, наблюдаемых в небольшом числе свободных от речи кадров. Недостатком подхода, основанного на классификации речевых параметров, является его слабость к шуму, которые имеют сходные спектральные характеристики с характеристиками речевого сигнала. В этом случае речевые параметры не смогут отличить речевые и шумовые компоненты доминирующей частоты, что приводит к низкой точности маскировки.

Список использованных источников:

1. Togneri R., Pullella D. An Overview of Speaker Identification Accuracy and robustness Issues // IEEE Circuits and systems magazine. 2011. P. 23–58.

2. M. Cooke, P. Green, and M. Crawford, “Handling missing data in speech recognition,” in Proc. IEEE Int. Conf. Acoustics Speech Signal Processing (ICASSP), 1994, pp. 1555–1558.

3. S. Boll, “Suppression of acoustic noise in speech using spectral subtraction,” IEEE Trans. Acoust. SpeechSignalProcess., vol. 27, no. 2, pp. 113–120, 1979.

4. M. El-Maliki and A. Drygajlo, “Missing features detection and handling for robust speaker verification,” in Proc. European Conf. Speech Communication Technology (Eurospeech), Budapest, Hungary, 1999, pp. 975–978.

5. J. Barker, M. Cooke, and P. Green, “Robust asr based on clean speech models: An evaluation of missing data,” in Proc. European Signal Process. Conf. (EUSIPCO), Aalborg, Denmark, 2001, pp. 213–216.

6. P. Janˇcoviˇc and M. Kцkьer, “Estimation of voicing-character of speech spectra based on spectral shape,” IEEE Signal Process. Lett., vol. 14, no. 1, pp. 66–69, 2007.

7. K. J. Palomdki, G. J. Brown, and D. Wang, “A binaural processor for missing data speech recognition in the presence of noise and small-room reverberation,” S peech Commun., vol. 43, no. 4, pp. 361–378, 2004.

8.Кручок, Д. Н. Эффект бинауральной маскировки для идентификации диктора в акустических шумах / Д.Н. Кручок // Современные технологии в науке и образовании – СТНО-2017 [текст]: сб. тр. междунар. науч.-техн. и науч.-метод. конф.: в 8 т. Т.3./ под общ. ред. О.В. Миловзорова. – Рязань: Рязан. гос. радиотехн. ун-т, 2017; Рязань. – 292 с. – С. 165–168.

9.B. Raj, “Reconstruction of incomplete spectrograms for robust speech recognition,” Ph.D. dissertation, Pittsburgh, PA, Carnegie Mellon Univ., 2000.

10. M. L. Seltzer, B. Raj, and R. M. Stern, “A Bayesian classifi er for spectrographic mask estimation for missing feature speech recognition,” Speech Commun., vol. 43, no. 4, pp. 379–393, 2004.

СТИЛИЗАЦИЯ ГОЛОСА С ИСПОЛЬЗОВАНИЕМ ГЛУБОКОГО ОБУЧЕНИЯ

Белорусский государственный университет информатики и радиоэлектроники
г. Минск, Республика Беларусь

Крылов Н.Д.

Одной из самых сложных задач для вычислительной техники является обработка естественного языка. В частности, одной из таких задач является имитация человеческого голоса. Алгоритмы машинного обучения являются наиболее эффективным инструментом в этой области.

Для обучения нейронной сети необходим правильно составленный набор данных. Будет ли правильным обучающим примером для данной задачи сказанная одна, двумя разными людьми, фраза? Фраза может быть произнесена с различной скоростью, громкостью и темпом, входная и выходная последовательность будет иметь различную длину. Обучение нейронной сети на таких данных затруднительно, в таком случае хорошим решением является использование двух нейронных сетей. Первая нейронная сеть используется в качестве кодировщика и переводит данные в иное представление, вторая в качестве декодировщика, позволяет получить требуемый результат.

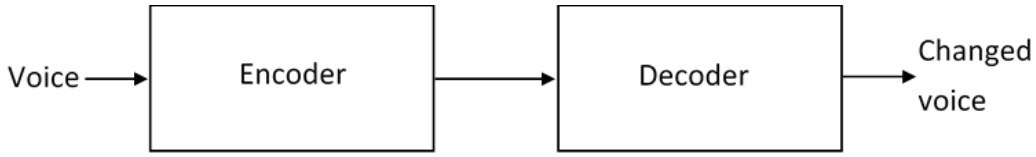

Рис. 1 – Общая структура модели

Такая модель успешно используется в задачах машинного перевода и подписи изображений. Для обучения такой модели необходим набор данных, содержащий пары звук-текст [1]. Обучение модели разделено на следующие шаги:

1. обучение системы распознавания речи (кодировщика), используя звук-текст обучающие примеры;
2. удаление Connection temporal classification и Softmax слоеввнейронноисети;
3. блокировка всех слоев кодирующей нейронной сети чтобы избежать их переобучения;
4. подключение к модели синтеза голоса (декодера);
5. обучение конечной модели с использованием только звука в качестве входных и выходных данных.

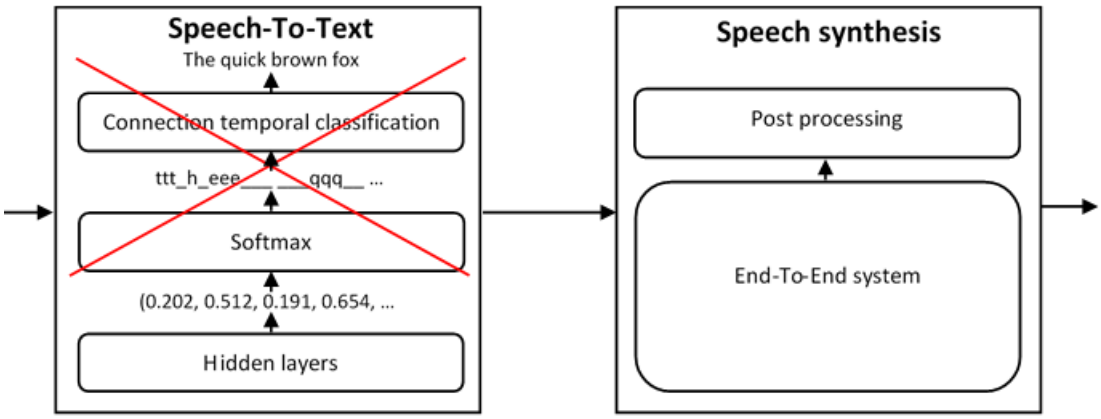

Рис. 2 – Подробная структура модели

В качестве кодировщика может быть использована предобученная нейронная сеть, такая как Speech-to-Text-WaveNet, DeepSpeech. Нейронная сеть для синтеза голоса должна быть обучена с нуля из-за нестандартного представления входных данных. Такое представление данных может содержать больше информации об интонации. Кодирующая нейронная сеть может быть использована для любых примеров, однако декодирующая должна быть дообучена в отдельности для синтеза конкретного голоса [2].

Список использованных источников:

1. AaronvandenOord, SanderDieleman, HeigaZen, KarenSimonyan, OriolVinyals, AlexGraves, NalKalchbrenner, AndrewSenior, KorayKavukcuoglu, WaveNet: AGenerativeModelforRawAudio [Электронный ресурс]. – 2016. – Режим доступа: https://arxiv.org/abs/1609.03499.

2. Wei Ping, Kainan Peng, Andrew Gibiansky, Sercan O. Arık, Ajay Kannan, Sharan Narang, Deep Voice 3: Scaling Text-to-Speech with Convolutional Sequence Learning [Электронныйресурс]. – 2018. – Режимдоступа: https://arxiv.org/abs/1710.07654.

ПОВЫШЕНИЕ КАЧЕСТВА ИЗОБРАЖЕНИЯ НА БАЗЕ АЛГОРИТМОВ НЕЙРОННЫХ СЕТЕЙ

Белорусский государственный университет информатики и радиоэлектроники
г. Минск, Республика Беларусь

Никитин Г.Ю.

Петровский А.А. – д.т.н., профессор

В настоящее время остро стоит проблема наличия различных шумов и артефактов на изображениях, что приводит к сильному снижению их качества. На практике наиболее распространённые шумы колеблются от аддитивного шума до мультипликативного шума. Такая деградация может оказать значительное влияние на точность методов компьютерной обработки в медицине, сделать анализ и распознавание изображений трудными и ненадёжными.

Существует большое множество алгоритмов, способных убрать с изображения шумы. В последнее время было проведено много исследований алгоритмов шумоподавления в вейвлет области, например, был предложен алгоритм на основе вейвлет преобразования с последующей отсечкой шумовых коэффициентов [1], или, к примеру, алгоритм оценки шумовых вейвлет коэффициентов, основывающийся на моделях