УДК 004.8

МЕТОДИКА ПЕРЕНОСА РЕЧЕВОГО АКЦЕНТА ДИКТОРА В СИСТЕМАХ СКВОЗНОГО ПЕРЕВОДА РЕЧИ НА ОСНОВЕ ДИСКРЕТНЫХ ПРЕДСТАВЛЕНИЙ

Островский А.О. [1,2]
Научный руководитель: д.т.н., профессор Карпов А.А.[1,2]
[1]Университет ИТМО, [2]СПб ФИЦ РАН
karpov@itmo.ru

Введение

Современные системы сквозного перевода речи (Speech-to-Speech Translation) [1] достигли значительных успехов как в точности передачи лингвистического содержания, так и в сохранении паралингвистических характеристик говорящего [2], таких как тембр голоса, темп речи, общая интонационная структура. Однако данные характеристики переносятся, как правило, неконтролируемо: система воспроизводит усреднённые акустические свойства, не предоставляя возможности целенаправленно управлять отдельными просодическими составляющими. В частности, акцент говорящего либо сглаживается до нейтрального, либо воспроизводится непредсказуемо в зависимости от обучающих данных. В данной работе мы сознательно концентрируемся именно на акценте как на одной из наиболее информативных и при этом явно детектируемых паралингвистических характеристик. Акцент несёт в себе информацию о языковом и культурном происхождении говорящего, играет важную роль в задачах дубляжа, однако задача явного переноса акцента в системах перевода речи остаётся слабо изученной [3]. Предлагаемый подход основан на разделении лингвистического содержания и просодического стиля с использованием дискретных акустических представлений.

Основная часть

Предлагаемый подход основан на парадигме разделения содержания и стиля (content-style disentanglement) [4]. Лингвистическое содержание речи представляется в виде последовательности дискретных единиц, получаемых посредством квантования скрытых состояний предобученной акустической модели. Такое представление, близкое к уровню фонем, очищено от информации о голосе и просодии говорящего [5]. Просодия, включающая акцент, кодируется отдельно через контур основного тона (F0) и характеристик и энергии сигнала. В качестве основного компонента системы используется кодер на базе рекуррентных сетей с управляемыми блоками (Gated Recurrent Units, GRU), обучаемый формировать усреднённое акцентное эмбеддинг-представление по группе дикторов с одинаковым акцентом.

Полученное представление конкатенируется с эмбеддингом клонированного голоса и передаётся в систему синтеза речи (Text to Speech, TTS). Подобный подход к построению пайплайна синтеза с явным управлением акцентом согласуется с архитектурными решениями, предложенными в работе Viglino et al. [6]. Эксперименты проводились на корпусе L2-ARCTIC [7], содержащем записи английской речи носителей различных языков (арабского, хинди, китайского, вьетнамского и др.). Дополнительно были собраны данные для французского и испанского языков. Система тестировалась в двух конфигурациях: перенос акцента между разными языковыми доменами (нигерийский и индийский акценты английского языка в направлении британского стандарта) и перенос внутри одного языка (канадский французский в направлении парижского стандарта).

Для оценки качества переноса использовалась модель Whisper [8]. Результат, наблюдаемый в том, что применение акцентного эмбеддинга меняет автоматически определяемый Whisper язык, позволяет использовать его в качестве косвенной метрики интенсивности переноса акцента: успешный перенос французского акцента на английскую речь приводит к тому, что модель определяет язык фрагмента как французский. Этот эффект, зафиксированный в экспериментах (например, клонирование голоса немецкоязычного диктора и последующий синтез испанской речи с добавлением французского акцентного эмбеддинга), показывает работоспособность предложенного подхода. Визуальный анализ спектрограмм и временных диаграмм также демонстрирует различимые изменения в частотно-временной структуре речи после добавления акцентного эмбеддинга при сохранении общей разборчивости.

Выводы

Представленная работа демонстрирует возможность переноса стиля акцента в системах синтеза и перевода речи посредством явного акцентного эмбеддинга. Разделение лингвистического содержания и просодического стиля через дискретные представления позволяет управлять акцентными характеристиками генерируемой речи без значительного ухудшения разборчивости. Предложенное использование модели Whisper как косвенной метрики интенсивности акцента открывает перспективы для разработки полностью автоматических методов оценки, не требующих разметки. Практическая область применения включает системы автоматического дублирования с сохранением идентичности говорящего, а также инструменты для обучения произношению. Направлением дальнейших исследований является распространение подхода на перенос эмоциональной окраски и темпоральных характеристик речи, а также разработка контролируемого механизма регулировки интенсивности акцента.

Литература

1. Lee A., Chen P.J., Wang C., Gu J., Popuri S., Ma X., Polyak A., Adi Y., He Q., Tang Y., Pino J., Hsu W.N. Direct speech-to-speech translation with discrete units // In Proc. of 60th Annual Meeting of the Association for Computational Linguistics (ACL). 2022. P. 3327–3339. https://doi.org/10.18653/v1/2022.acl-long.235
2. Popuri S., Chen P.-J., Wang C., Pino J., Adi Y., Gu J., Hsu W.-N., Lee A. Enhanced direct speech-to-speech translation using self-supervised pre-training and data augmentation // In Proc. of Interspeech 2022. 2022. P. 5195–5199.
3. Grinstein E., Duong N.Q.K., Ozerov A., Perez P. Audio style transfer // In Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2018. P. 586–590. https://doi.org/10.1109/ICASSP.2018.8461711
4. Nguyen T. N., Pham N.Q., Waibel A. Accent conversion using discrete units with parallel data synthesized from controllable accented TTS // arXiv:2410.03734. 2024.
5. Hsu W. N., Bolte B., Tsai Y. H.H., Lakhotia K., Salakhutdinov R., Mohamed A. HuBERT: Self-supervised speech representation learning by masked prediction of hidden units // IEEE/ACM Transactions on Audio, Speech, and Language Processing. 2021. Vol. 29. P. 3451–3460. https://doi.org/10.1109/TASLP.2021.3122291
6. Viglino T., Motlicek P., Cernak M. End-to-end accented speech recognition // In Proc. of Interspeech 2019. 2019. P. 2140–2144. https://doi.org/10.21437/Interspeech.2019-2122
7. Zhao G., Sonsaat S., Silpachai A., Lucic I., Chukharev-Hudilainen E., Levis J., Gutierrez-Osuna R. L2-ARCTIC: A non-native English speech corpus // In Proc. of Interspeech 2018. 2018. P. 2783–2787. https://doi.org/10.21437/Interspeech.2018-1110
8. Radford A., Kim J.W., Xu T., Brockman G., McLeavey C., Sutskever I. Robust speech recognition via large-scale weak supervision // In Proc. of the 40th International Conference on Machine Learning (ICML). 2023.Vol. 202. P. 28492-28518.